

Tribal culture versus genetics

A dispute between researchers and a small Native American tribe has cast an unduly large shadow over genetics. Both sides have much to gain from deeper communication, aided by those who belong to both communities.

Native Americans in the southwestern United States are in conflict with researchers over a genetics study. The Havasupai tribe has engaged scientists and universities in a legal battle over a diabetes research project. The tribe's lawsuits allege that researchers from Arizona State University and other institutions used DNA for studies without proper consent. The project's leader emphatically denies this, and researchers are stunned at the allegations about their (as they see them) benevolent efforts.

Politically charged interpretations of the dispute have spread through Native American communities, fanning tribal distrust of academics. Ill-founded rumours are the last thing that tribes and scientists need at this promising time for genomic research.

Such studies last for years, and the dispute highlights the importance of researchers keeping in constant communication with their Native American research subjects throughout (see page 500). History has shown that a court of law is not a good forum for resolving ethical debates. But how to balance the need for sensitivity to tribal culture while fostering rigorous scientific inquiry?

Some tribes now maintain their own human-subjects committees, which must approve all research projects. There are advisory committees to assist groups in monitoring projects whose complexities are difficult for non-scientists to understand. And there are proposals for Native American gene banks controlled and monitored by the tribes themselves — a concept that could provide them with ownership of products that may be derived from their genes.

All of these initiatives offer promising opportunities, but they also come with responsibilities for both researcher and subject. If subjects want to know what the researchers are trying to accomplish, and are kept informed, clashes of culture and science may be prevented.

Moreover, there have been a growing number of instances in which repeat consent was sought from research subjects from special communities, where language and cultural barriers may complicate projects. While worthy, this concept — time-consuming and potentially costly — would be unnecessary if there was a regular two-way flow of information between researchers and subjects.

Some ethicists suggest that an obsession with the details of consent have caused research subjects to forget they have an opportunity to help not only their own tribe, but all mankind. For Native Americans, this is a hard concept to accept. Having seen their people and cultures abused for centuries, they are understandably hypersensitive. But it could be a new form of empowerment for them to realize that their culture helped cure a disease.

Today, many Native American tribes have economic opportunities they never dreamed of, including education and access to scholarships. Gaming revenues provide better community services and chances to eliminate the sicknesses of poverty that for generations have plagued reservations. But too often this new-found economic clout is used to further litigation for tribal political purposes. In Arizona, sensitive, caring scientists are privately saying they do not want to go anywhere near a reservation after recent events in the Havasupai case. Given the broader potential benefits of research, this cannot be a climate that tribes wish to foster.

Leaders from both communities need to reach out to each other to bridge the gap between their cultures. The National Human Genome Research Institute is funding work to do precisely this. One group in a unique position to help are Native American scientists: they too can support dialogues to create a research environment to match the genetic opportunities of the times. ■

States versus gases

A state-led lawsuit against greenhouse-gas emitters highlights a forceful regional movement in US climate policy.

If you had to predict who would save the world, city lawyers may not instantly spring to mind. So many people were surprised last week when US lawyers launched a strike against global warming.

Attorney-generals from eight states and lawyers from New York City filed a lawsuit demanding cuts in emissions from the five major power companies that they say belch out about 10% of the nation's carbon dioxide (see www.nature.com/news/2004/040719/full/040719-12.html). The move is an unmistakable dig at the Bush administration for shirking strict curbs on greenhouse-gas emissions in favour of voluntary reductions.

Much of the lawsuit is sheer showmanship from the ambitious legal team behind it. When they get their day in court, they may struggle to win a guilty verdict. It may be tough to prove that a few companies should shoulder the blame for their share in a global problem, or that the modest cuts the lawsuit seeks would help.

But the trial signals that the fight against global warming in the United States is far from over. Lawyers and policy-makers in individual states are willing to take the issue into their own hands — even if

President George Bush is sitting on his. And the states can force the federal government to deal with issues where activists have failed.

The Regional Greenhouse Gas Initiative, for example, is an effort by nine northeastern and mid-Atlantic states to build a system to cap greenhouse-gas emissions. California, meanwhile, is developing legislation demanding curbs in carbon dioxide from vehicles. Such initiatives could drive a change in national policy simply by showing that it can be done, or because companies reined in by conflicting state laws may turn to the federal government for clarity.

Many environmentalists would like to see the heads of power plants squirm in the dock, and may get their wish. Some experts predict an imminent wave of lawsuits against greenhouse-gas producers, much like those against the tobacco industry, from people claiming damages for property or loved ones lost to floods or droughts.

But acrimonious court battles are not the best way to resolve issues that affect the future of the planet. Negotiation, legislation and regulation are. State lawmakers should unite and act where the federal government has not; scientists and activists should support them. ■





High stakes
Researchers launch
lawsuit to secure
legal marijuana
p492



Whisked away
Dinosaur eggs saved
from the clutches of
private collector
p493



Out of the House
Congressional
firebrand heads
for industry
p495



Making a splash
Improved cash flow
aids restoration of
Iraq's wetlands
p496

Lawyers blast nuclear pact as a breach of disarmament treaty

Jim Giles, London

The impending renewal of a pact on nuclear research between the United Kingdom and the United States could breach the Nuclear Non-Proliferation Treaty (NPT), British lawyers say.

Critics argue that the two countries have long been in violation of the NPT, the cornerstone of international attempts to halt the spread of nuclear arms, both in spirit and in the letter of the law. But this year's pending renewal of the US/UK Mutual Defence Agreement (MDA) prompted advocacy groups to seek a legal opinion on the matter. Armed with this advice, they are hoping to force both countries to take disarmament more seriously.

The MDA dates from 1958 and allows for the exchange of plans for the research and development of nuclear weapons between Britain and the United States. Such collaboration aided the development of Britain's nuclear weapon system, Trident, for example. But both countries have also signed the NPT, which came into force in 1970 and requires them to work towards disarmament.

Britain and the United States have improved their weapons systems since the NPT became binding, arguing that they have stuck to its spirit by working to reduce the overall number of nuclear arms. Disarm-



Collaboration between US and UK scientists helped develop the Trident nuclear missiles.

ament advocates counter that any work to improve weapons is a legal breach of the treaty.

It is a grey area, but lawyers have now tipped the balance in favour of those who criticize the MDA. "It is strongly arguable that the renewal of the MDA is in breach of the NPT," conclude Rabinger Singh and Christine Chinkin, lawyers at Matrix Chambers in London, who were hired by a number of UK-based advocacy groups.

Neither Britain nor the United States will discuss details of the research that goes on

under the agreement. A spokesman for the UK Ministry of Defence points out that the allies regularly exchange information on how to safeguard existing nuclear weapons, rather than information about developing new ones. But researchers who took part in previous MDA collaborations, including Bob Peurifoy, former chief weapons designer at Sandia National Laboratories in Albuquerque, New Mexico, say that weapons development has definitely been part of discussions in the past. "The designs were better because of the exchange," Peurifoy says.

Britain has already declared its intention to renew the MDA for a further 10 years, and the agreement is expected to pass through US Congress without debate.

The British American Security Information Council, one of the advocacy groups that commissioned the Matrix lawyers, says that it is considering seeking a judicial review of the UK decision to renew the MDA — if it can come up with the money needed to do so.

Daryl Kimball, executive director of the Arms Control Association in Washington DC, says the lawyers' conclusion highlights the secrecy that surrounds the MDA. "The United States and the United Kingdom should be more transparent. If no there is no violation, they should provide more information," he says. ■

Energy labs halt classified research amid security fears

Geoff Brumfiel, Washington

The bulk of classified research at all 24 labs run by the US Department of Energy was shut down this week after officials decided that the security problems recently found at the Los Alamos National Laboratory in New Mexico might also exist elsewhere.

Several computer disks went missing at Los Alamos earlier this month, prompting energy secretary Spencer Abraham to halt all work at the lab until the problems are ironed out (see *Nature* 430, 387; 2004). As of 26 July, two missing disks had still not been found, and 15 employees had been suspended in

connection with the disappearance.

Now all energy labs will stop doing classified research that involves removable storage devices — such as computer disks — until all the devices are accounted for and new procedures are in place for monitoring their handling by laboratory employees.

"While we have no evidence that the problems currently being investigated are present elsewhere, we have a responsibility to take all necessary action to prevent such problems," Abraham said on 23 July.

The shut-down isn't quite as dramatic as it sounds, experts say. Only two labs will

be seriously affected: Sandia National Laboratories in Albuquerque, New Mexico, and Lawrence Livermore National Laboratory in California. Together with Los Alamos, these two labs conduct the bulk of the country's nuclear-weapons programmes. At Livermore, 876 employees will be suspended during the inventory of some 12,000 items of classified removable material.

At other labs, far fewer people will be affected. "The impact should be minimal," says Martha Krebs, former director of the energy department's science office. ■

Joint suits aim to weed out agencies' red tape

Helen Pearson, New York

Frustrated with the slow pace of research into medicinal marijuana, researchers have launched a pair of lawsuits accusing US government agencies of obstructing attempts to obtain supplies of the plant.

In order to study marijuana for its ability to ease pain, nausea or symptoms of AIDS, US researchers procure the drug from a small farm at the University of Mississippi, under contract from the National Institute on Drug Abuse (NIDA). But some complain that the red tape and long delays involved in getting the plant through NIDA, the Drug Enforcement Administration (DEA) and other agencies are unacceptable.

The lawsuits, filed on 22 July, were coordinated by the Florida-based campaign group Multidisciplinary Association for Psychedelic Studies (MAPS).

In one lawsuit, MAPS is demanding a decision on an application filed in 2003 to buy 10 grams of NIDA marijuana — a tiny amount worth just \$70, the group says. The researchers plan to use the plant in testing a vaporizer, an alternative method of delivery to smoking. The test would look at the chemical constituents of the vapour and would not involve human subjects.

The second lawsuit contends that the DEA has stalled an application to set up a separate farm to grow marijuana for research, filed in June 2001 by botanist Lyle Craker of the University of Massachusetts at Amherst. If successful, this would be



Strong blow: federal agencies are being sued for hampering researchers' access to marijuana.

the first official alternative US marijuana source for medical researchers.

MAPS president Rick Doblin argues that the NIDA supply of marijuana is of low quality and potency, and that researchers will be unable to get clinical approval for drugs derived from it unless they can grow a pharmaceutical-grade crop themselves. He would

like to see a situation like that in Britain, where Salisbury-based GW Pharmaceuticals, under licence from the Home Office, has established its own greenhouse facility to grow marijuana for clinical trials.

The US Department of Health and Human Services, NIDA and the DEA would not comment on the lawsuits. But those behind the legal action claim that the agencies are sitting on the applications because they go against the federal government's hard line on drugs. "It's politically unacceptable to say yes," says Craker (see *Nature* **430**, 394–395; 2004).

Other researchers in the field agree that the bureaucratic application process for marijuana studies has contributed to sluggishness in the field. "The pace has been slower than one would like," says mental-health researcher Stanley Watson of the University of Michigan, Ann Arbor, who co-authored a 1999 Institute of Medicine report urging clinical trials into medicinal marijuana.

But that does not mean that NIDA is doing anything illegal, Watson says. The agency's remit is to study research linked to the abuse of drugs, he points out, not their medicinal use. "It's going to be a tough one for MAPS to win," says Watson. Others add that there could be scientific problems with the MAPS applications that are holding things up.

Watson suggests that researchers might best forge ahead by improving the design of their trials, or perhaps by negotiating marijuana supplies from other countries. ■

Sea snapshots will map frequency of freak waves

Michael Hopkin, London

They are known as 'rogue waves' — the towering walls of water that, some experts suspect, sink tens of ships every year. Now oceanographers are planning to use satellite images to produce a global map of where and how often the rogues occur.

The WaveAtlas initiative follows a trial using three weeks' worth of radar images obtained by the European Space Agency (ESA) from its two European Remote Sensing (ERS) satellites. The trial data covered February and March 2001, a period during which two tourist liners, the *Bremen* and the *Caledonian*, had their windows smashed by 30-metre waves in separate incidents in the Southern Ocean.

The trial's results make hair-raising reading. Besides the two 30-metre giants, the team identified at least eight other waves topping 25 metres across the world. It is a wake-up call for anyone who views rogue waves as a nautical myth, says project

member Wolfgang Rosenthal of the GKSS Research Centre in Geesthacht, Germany.

The full project, which will encompass two years' worth of images from 1998 to 2000, could help to explain the staggering number of unexplained sinkings worldwide. "There are many more accidents than you would think — around two a week," says Rosenthal. "They simply get put down to bad weather."

As the ERS satellites circle the globe, they each take a representative radar snapshot of an area 10 kilometres by 5 kilometres for every 200 kilometres of the Earth's surface they cover. The WaveAtlas team uses the amount of radar reflected to determine the incline of the ocean surface, and therefore the size of the waves captured in the image. The researchers have received 75% of the requested images from ESA, and hope to complete the analysis early next year.

Project leader Susanne Lehner, a marine physicist at the University of Miami, Florida,



Surf's up: an artist's impression of a rogue wave.

suspects that rogue waves can form when existing waves are chased by a storm system moving at roughly the same speed. "We want to see what weather patterns they are associated with," she says.

The project could also inform the design of ships and oil platforms, Rosenthal argues. Most current platforms are built with a clearance of 15 metres, he says. "The designers think they did a good job, but officers on the platform say 'we get wet feet'." ■



Bad eggs: a New York auction house decided not to put fossilized embryos under the hammer.

Dinosaur eggs escape sale as smuggling claims unearthed

Rex Dalton, San Diego

An outcry among palaeontologists — and a little help from US federal agents — has saved some dinosaur embryo fossils from disappearing into the living room of a wealthy patron. But questions remain as to how the half-dozen specimens, thought to have been smuggled out of Argentina, will be repatriated for study.

The fossils, which were set to be auctioned on 24 June at Guernsey's in New York City, include some valuable specimens, such as a dinosaur egg in which the skull of a sauropod embryo can be seen. "It is a tremendous specimen, extremely rare," says Luis Chiappe, an Argentinian palaeontologist who is a curator at the Natural History Museum of Los Angeles County. Such detailed fossils could be used to study the early development of dinosaurs.

When Argentinian government officials learned from US palaeontologists that the illicit dinosaur embryos were to be sold, they contacted the US government, which took the unusual step of directing the Federal Bureau of Investigation to contact Guernsey's auction house. The embryo fossils were then removed from the auction. They are now being held in New York, and efforts are under way to get them back to Argentina.

Palaeontologists have also raised questions about the background of other fossils that went unsold in the auction, which were listed as being from China but which scientists say were probably smuggled from Mongolia. These fossils are still owned by commercial fossil-dealer Zee Haag of Tucson, Arizona.

The Argentinian fossils were brought to the auction by Terry Manning of Leicester, UK, who has collected dinosaur eggs for about 15 years. He has refined an acid-etching technique to reveal embryonic material in fossils.

Manning says he bought the dinosaur eggshells two years ago at a fossil show in Tucson, Arizona. He says he took the shells to England to clean them up, and then brought them to New York, mentioning to US customs agents that the specimens were probably originally smuggled from Argentina. The agents told him to "Have a good auction," he says.

Although Manning says he would appreciate some money for his fossils, he adds: "The most important thing is that the research work must be completed on the specimens." This is only likely to happen if the fossils wind up in government hands, rather than in a private collection. The Argentinian government is keen to take them, although Marcelo Cema, a spokesman at the Embassy of the Argentine Republic in Washington, is adamant that it will not pay for transportation of the "stolen property".

If anyone wants to retrieve Haag's fossils to send them home to Asia, it will cost them. Haag says his three-metre-long *Tyrannosaurus bataar* skull, for example, is worth at least US\$160,000. Palaeontologist Mark Norell of the American Museum of Natural History, New York, says someone probably smuggled the skull out of Mongolia. Haag says he bought the skull and other specimens in Tokyo, and shipped them to Arizona, claiming them properly for import. As to whether these specimens would be better off in a lab than on the auction block, Haag simply says: "I'm an American capitalist." ■

Swedish enthusiasm peps up plans for neutron source

Quirin Schiermeier, Munich

An ambitious plan for a powerful neutron facility in Europe has been reawakened by a show of interest from Sweden. There are as yet no promises that the facility will be built — let alone a decision on where — but Sweden's move has fuelled hope in the 5,000-strong European neutron community that the project will finally go ahead.

Europe announced plans in 1992 to build the European Spallation Source (ESS), a facility designed to be the most powerful neutron source in the world. The ESS would produce neutrons by accelerating protons at a heavy metal target — such neutrons can be used to probe materials from proteins to plastic and steel. Although other neutron sources are planned in the United States and Japan, the ESS would be more powerful and flexible than these.

By 2002 the technical plans were finalized, but in 2003 both Germany and Britain withdrew support, in a move that seemed to kill off the €1.5-billion (US\$1.8-billion) project (see *Nature* 421, 563; 2003).

On 16 July, the Swedish government asked former minister of finance, Allan Larsson, to review the possibility of hosting the facility. Larsson has a one-year mandate to garner support for the project — or a scaled-down version — from European governments, science agencies and industry. If he succeeds, next summer Sweden will submit a formal bid.

"I am absolutely delighted," says Bob Cywinski, a physicist at the University of Leeds, UK, and a long-time lobbyist for a Yorkshire site for the ESS. "Of course I'd favour it being built in Britain, but Sweden's move will definitely dispel the notion that the ESS is a dead project." At the same time, an umbrella organization — the European Spallation Source Initiative — is being set up in Grenoble, France, to oversee the project.

There are other contenders for hosting the site, however. Hungary is expected to come up with a formal bid later this year. And Britain, which operates what is currently the world's most intense neutron source at the Rutherford Appleton Laboratory near Oxford, may also remain an option.

It would take about ten years to build the ESS, so advocates are pressing for a decision to be made on its location as soon as possible. "We could start building tomorrow," says Cywinski. ■

Winged messenger set to follow ancient mariner to Mercury

Tony Reichhardt, Washington

Scientists and engineers will hold their breath next week as NASA prepares to shoot a spacecraft towards one of the last unexplored planets of our Solar System — Mercury.

The Mercury mission had significant challenges to overcome before reaching the launch pad. For example, it had to be designed to withstand temperatures of more than 400 °C and to perform a tricky orbital insertion. At the time it was approved in 2001, NASA officials called it the most complex Discovery-class planetary mission they had ever attempted. Discovery missions are designed to have the relatively low cost of less than US\$400 million.

The Mercury Surface, Space Environment, Geochemistry and Ranging (MESSENGER) spacecraft is set to leave Cape Canaveral, Florida, on 2 August, with launch windows every day until 14 August. It should arrive at Mercury in March 2011.

The only other visitor to Mercury has been Mariner 10, which mapped 45% of the planet's rocky, cratered surface in 1974 and 1975. Mariner left key questions unanswered, such as whether the surface rocks are volcanic in origin, and why Earth and Mercury have global magnetic fields, but Mars and Venus do not. A suite of seven instruments onboard MESSENGER, including X-ray, γ -ray and infrared spectrometers, will try to come up with answers, while cameras will map the entire surface at high resolution.

Getting MESSENGER into orbit around a planet so small and so close to the Sun will not be easy. In fact, it was considered impractical using current rockets until a mission designer in the 1980s found a way to slow down a spacecraft using a complicated series of loops around Venus and Mercury.

Even then the project nearly never took off. Last year NASA threatened to terminate it for running \$26 million over its agreed cost limit. The agency's Space Science Advisory Committee saved the project by advising NASA to spend the extra money, even if it meant delaying other planetary missions.

If successful, it will be a relief for mission designers at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland. The lab's last Discovery spacecraft built for NASA — the Comet Nucleus Tour (CONTOUR) — failed shortly after leaving Earth in 2002. ■

► <http://messenger.jhuapl.edu>

Russian bid to drill Antarctic lake gets chilly response

Jim Giles

Russian researchers are preparing to drill into an underground lake in Antarctica despite concerns from fellow polar scientists that the plan may be too hasty.

The Russian team will begin drilling down to Lake Vostok, which lies nearly 4,000 metres below the surface and is thought to have been isolated for 20 million years, in December.

But experts say that the team's equipment has not been properly tested and could contaminate the potentially unique ecosystems that inhabit Vostok's 5,400 cubic kilometres of water.

"The lake is too valuable to experiment with," warns Cynan Ellis-Evans, a microbiologist at the British Antarctic Survey in Cambridge, UK. Researchers hope that any organisms found in the lake could shed light on both early life on Earth and the possibility of life existing on other icy planets.

The Russian plans are likely to spark heated debate at a meeting of the Scientific Committee on Antarctic Research, held in Bremen, Germany, on 25–31 July. Many of the Antarctic scientists contacted by *Nature* before the meeting suggested that the researchers should drill into one of the more than 100 other sub-glacial lakes instead — both to test the equipment and to study the ecosystems.

The other lakes may not be as old as Vostok, but their smaller size may ease sampling of their waters, says Martin Siegert, a glaciologist at the University of Bristol, UK. "There is huge kudos associated with the first exploration of Vostok," he adds. "It's very difficult to get another lake on the agenda."

The Russian team, led by Valery Lukin of the Arctic and Antarctic Research Institute in St Petersburg, plans to extend a 3.5-kilometre-deep borehole that already exists at the Vostok station. Work on that hole stopped in 1998 when it was realized that unsterilized drilling equipment was just 130 metres from the lake's surface. The Russians hope to break through to the lake in 2006–07.

Water from the lake, which is at high pressure, should move up into the borehole and freeze. Lukin's team will then drill into and extract this frozen water. The group says that its drill is designed to prevent any surface contamination from entering the lake.

But polar scientists say that the Russian plans, while reasonable on paper, need to be



Russian researchers hold a core sample taken from above Lake Vostok. They say they will soon access the lake itself.

tested. They point out that the borehole contains about 60 cubic metres of kerosene, added to prevent the drill freezing, which is contaminated with surface microbes. And the lake water, which is at about 400 atmospheres, could explode into the borehole. "It will be like drilling through the cork of a champagne bottle," says Ellis-Evans.

Experts say that the Russian researchers have addressed some of the international community's concerns. But "they have dealt poorly with the contamination issues," says Jean-Robert Petit, an Earth scientist at the Laboratory of Glaciology and Environmental Geophysics near Grenoble, France, and a member of an international team developing longer-term plans to enter Lake Vostok. As no one country has control over Antarctica, Petit and others are powerless to prevent the drilling. "There is no law to stop them doing what they want," he says.

Researchers familiar with the Russian programme say privately that the plans may have a political motive. They suggest that Lukin and his colleagues were forced to support a national project in order to get funding from the Russian government, as an international effort would have less prestige in the government's eyes. These experts hope that the Russian team will use the plan to generate funding but won't actually enter the lake. But Lukin insists that his team will sample the lake as planned. ■

Tough talker quits Congress for bioindustry

Meredith Wadman, Washington

A congressman fêted for his dogged pursuit of corporate wrongdoers has made a career change that some observers are finding hard to swallow. James Greenwood (Republican, Pennsylvania), who led recent attacks on conflict-of-interest policies at the US National Institutes of Health (NIH), is to become president of the biotechnology industry's major trade group.

Greenwood accepted the \$800,000-per-year job with the Biotechnology Industry Organization (BIO) on 21 July. He will complete his current congressional term and begin with BIO in January 2005.

"This whole thing makes a mockery of the oversight responsibility of Congress," says Vera Hassner Sharav of the Alliance for Human Research Protection in New York. "It invites industry to buy their congressmen when they don't like where they're heading," she says. Members of Congress make \$158,000 — less than a quarter of Greenwood's new salary.

Others insist that Greenwood has not been bought. "It's hard to believe he's for sale," says Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania, who knows Greenwood and has testified before him on stem-cell research. But he concedes that "there is something unnerving when the watchdog becomes the guard dog".

Caplan speculates that there may be a political motive underlying Greenwood's move. The increasingly right-wing House of Representatives may have become a tire-



All change: James Greenwood is drawing a line under his congressional career.

some, hostile environment for a more liberal Republican such as Greenwood, he says. Greenwood has supported abortion rights, generous budget increases for the NIH and liberal policies on stem-cell research.

Since accepting the BIO job, Greenwood has excused himself from affairs that might conflict with his future work at BIO. Key among them is a hearing — abruptly cancelled on 20 July and now rescheduled for 9 September — at which big pharmaceutical companies are to be quizzed about their alleged failure to disclose negative results from trials of antidepressants in children (see *Nature* 429, 589; 2004). Among the expected witnesses are top officials from Pfizer, Wyeth,

Bristol-Myers Squibb and Eli Lilly — all of whom are members of BIO.

The hearing, which was to have been chaired by Greenwood, will now be chaired by another congressman. This has left some worried that without Greenwood's tough questioning the session won't be as thorough as first imagined. "This was to be a grilling," says Sharav. "What kind of hearing are we going to have now?"

Greenwood was not available for comment to *Nature*, but he released the following statement about his new job. "I passionately believe in the promise of biotechnology to find cures and treatments for the diseases that force parents to watch their children suffer and die." ■

Biologists lobby China's government for funding reform

David Cyranoski, Beijing

A group of prominent US-based Chinese scientists met with a high-ranking official from China's government last week to complain about the country's biased and inefficient system for funding life sciences. China is currently finalizing plans for several massive 15-year projects, but the researchers are concerned that not all of these are warranted and that they are absorbing too much of the science and technology budget.

Much of China's research is funded by the Ministry of Science and Technology, but for years the system has been criticized for its inefficiency and lack of transparency.

These criticisms came to a head following China's poor response to the outbreak of severe acute respiratory syndrome (SARS) in late 2002. As a result, a number of researchers both inside and outside the country called for the government to set up a Chinese

equivalent of the US National Institutes of Health to guide funding in the life sciences (see *Nature* 428, 679; 2004).

But last week the science ministry rejected this proposal. Although researchers had worried that the plan would increase China's heavy bureaucracy, its rejection means that the funding system is still in desperate need of reform, they say. "It can't deal with conflicts of interest," says Haifan Lin, a stem-cell biologist at Duke University in Durham, North Carolina. "It's not a valid system."

At last week's meeting, Lin and other members of the Ray Wu Society presented a letter to a senior adviser of Wen Jiabao, the prime minister, signed by 11 members of the society's board and outlining their concerns. The non-profit society, named after a plant biologist and composed of senior Chinese scientists based mainly in the United States, aims to develop the life sciences in China.

They say that their worries have been intensified by the science ministry's plans to assign funding to large projects covering 2006–20 by early next year. "The Chinese system encourages the control of large sums of money by a few people, without fair competition," says Tian Xu, a geneticist at Yale University and a member of the Wu society's board.

Xu and his society colleagues say that the country's other main funding organization, the National Science Foundation of China, does review projects fairly. But they point out that its annual budget of 2.2 billion renminbi (US\$266 million) is small compared with that of the ministry.

Over the next few weeks, Lin and his colleagues will try to ensure that their plea for a better funding mechanism reaches the prime minister. "We are approaching the government at the highest level," he says. ■

Cuts to NASA's budget leave plans for Mars mission grounded

Washington Mars seems a little farther away now that the House of Representatives committee that determines NASA's budget has cut \$1.1 billion from the space agency's \$16.2-billion request for next year. The move is a blow to President George Bush's plan to send astronauts to the Moon and Mars (see *Nature* 427, 183; 2004).

On 22 July, the committee deleted \$438 million proposed for the Crew Exploration Vehicle, the first big hardware purchase intended for the programme. It also cut \$230 million from a nuclear-rocket development effort that figures prominently in the Moon-Mars plans and a scientific mission to Jupiter and its icy moons.

House majority leader Tom DeLay, a powerful Republican whose Texas district includes NASA's Johnson Space Center, branded the action "unacceptable" and vowed to oppose the cuts. The Senate is expected to weigh in on the issue in September, but the final decision is unlikely to be made until after the presidential election in November.



Water world: the marshes of Mesopotamia are being restored, creating a habitat for wild birds.

Cash floods in to save Iraq's Garden of Eden

Tokyo There is fresh hope for the damaged wetlands in Iraq that are thought to be the location of the biblical Garden of Eden.

The marshlands of Mesopotamia were reduced to less than a tenth of their original area of 20,000 square kilometres when the Tigris and Euphrates river systems were dammed and drained by Saddam Hussein's regime. But the Japanese government pledged on 23 July to invest US\$11 million to restore and protect the marshes, under a scheme run by the United Nations Environment Programme (UNEP).

Local residents have re-flooded a fifth of the wetlands since the collapse of Saddam's regime, although a lack of sanitation services led to the spread of water-borne diseases.

Blue whale makes a splash on return to Alaska

San Francisco Blue whales have been sighted for the first time in 30 years in the waters off Alaska, where they had been hunted close to extinction.

Researchers aboard the research vessel *McArthur II* photographed the blue whales (*Balaenoptera musculus*) some 200 kilometres southeast of Prince

William Sound. An ocean-floor recording system had picked up the sounds of blue whales communicating over the past few years, but the photographs, taken earlier this month, are the first to document their presence.

The cruise's chief scientist, Jay Barlow of the US Southwest Fisheries Science Center in



La Jolla, California, says the sighting was a bonus as his team had set out to study humpback whales. One of the photographed blue whales (seen here) appears to be the same individual that has been spotted in recent years off southern California, although further analysis of the photographs is needed to confirm this.

The UNEP project will equip about a dozen settlements with small-scale water treatment systems, and local people will be trained in wetland management. Reed beds, which act as natural water-filtration systems, will also be restored, providing habitats for birds and other wildlife.

Race is key to rapid approval of heart drug

New York A heart drug being tested in African Americans is on course to become the first medicine approved for use in a specific ethnic group.

A clinical trial of BiDil, developed by NitroMed of Lexington, Massachusetts, was stopped on 19 July because it seems to be so effective when used in addition to normal therapy. The drug, which relaxes blood vessels and eases the strain on the pumping heart, could be launched in 2005, if it is approved by the US Food and Drug Administration.

But the trial has stirred controversy. Some scientists say it would be better to search for the specific genetic variations involved in a response to a drug, rather than relying on race. Anne Taylor of the University of Minnesota, Minneapolis, who led the trial, says the team plans to scan the genes of those patients who responded to BiDil. In the meantime, she argues, race may serve as a reasonable surrogate when making prescriptions.

Bush backs biotech to build defence stockpile

Washington Biotech companies in the United States will receive billions of dollars of government support over the next decade under plans to tackle bioterrorism.

Project Bioshield, which was signed into law by President George Bush on 21 July,

sets aside \$5.6 billion for the government to purchase vaccines and antidotes to potential bioweapons from private firms. The legislation also contains provisions that will allow the health secretary, Tommy Thompson, to speed grants for research projects through "expedited peer-review procedures" if they are deemed essential to biodefence needs.

"Private industry plays a vital role in our biodefence efforts," Bush said at a signing ceremony at the White House. "By acting as a willing buyer for the best new medical technologies, the government ensures that our drug stockpile remains safe, effective and advanced."

Plankton provide route to monitoring climate

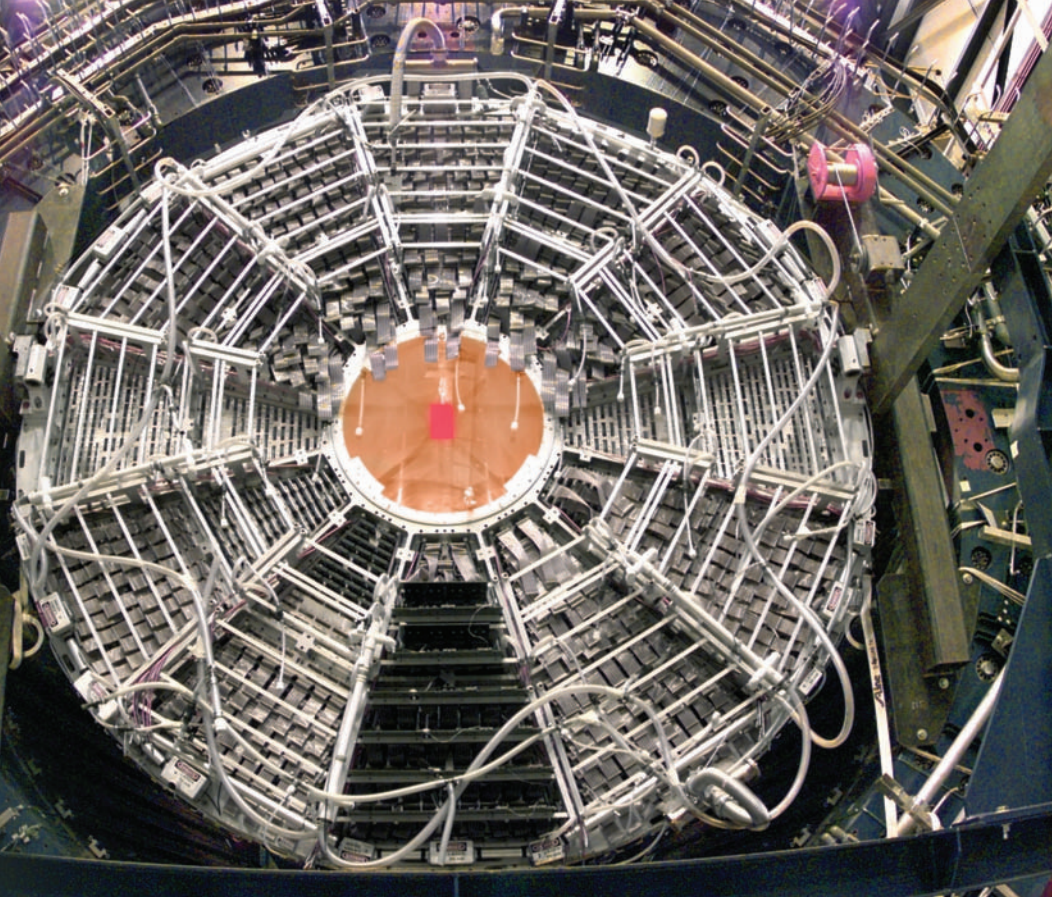
London Plankton could serve as an index for the effect of climate change on the world's aquatic systems, marine scientists have suggested.

The movement of ocean currents as a result of climate change is already known to have hit cod stocks. The warm-water phytoplankton on which they feed have moved 1,000 km northward in the eastern Atlantic over the past 40 years (G. Beaugrand *et al. Science* 296, 1692–1694; 2004). So tracking the distribution of plankton could help scientists to keep tabs on climate change, according to Chris Reid, director of the Sir Alister Hardy Foundation for Ocean Science in Plymouth, UK.

"Observing changes in plankton could provide an early warning for changes in marine and freshwater systems," agrees Martin Attrill, a marine ecologist at the University of Plymouth who organized and hosted the conference at which Reid made his proposal last week. Reid says that more plankton data will be needed to create the index, particularly from little-studied areas such as the Norwegian and Barents seas.

What's in a name?

Physicists agree that experiments at the Brookhaven atom collider have created a new form of matter. But theorists and experimentalists are still arguing about what to call it. Geoff Brumfiel investigates.



Just over a year ago, a rare event took place at Brookhaven National Laboratory on Long Island, New York. It didn't happen in the lab's massive atom smasher, but in a packed seminar room. Here experimentalists and theorists collided in a very public manner at a special colloquium held to announce "exciting new results".

Since June 2000, Brookhaven's particle accelerator, the Relativistic Heavy Ion Collider (RHIC), has smashed atoms together in a bid to make an extremely hot state of matter with a name only a physicist could love: the quark-gluon plasma. For decades, theorists have been creating detailed models of this plasma, which was thought to exist in the early Universe, and RHIC was built in an attempt to confirm or refute these calculations.

A year after the machine was switched on, the four detectors that make up RHIC saw signs of the new plasma state. Subsequent experiments re-examined those hints and generated a lot more data. So expectations among the physicists, politicians and journalists were high at the colloquium in June 2003. Were the experimental groups ready to announce the discovery of the quark-gluon plasma?

The announcement never came. Instead, researchers presented a series of long lectures on the physics happening

inside the collider. Detail by aching detail, members of the four detector teams laid out the case that RHIC had produced a very hot, dense form of matter that had never been seen before.

The reporters and dignitaries scratched their heads as they left the lecture hall. They had heard enough physics in two hours to give many of them a low-grade migraine, but they were no closer to knowing whether RHIC had created the strange state of matter it was built to discover.

Then came the press briefing. Top experimentalists and leading theorists gathered in another packed room to try to explain in simpler terms what the experiment had achieved. William Zajc, an experimentalist from Columbia University, New York, who represented RHIC's PHENIX detector, said that there might be a quark-gluon plasma in the machine. But he added: "I think we would like to rule out any more ordinary explanations first." Then Miklos Gyulassy, a Columbia theorist, gave everyone the answer they had been waiting for. "It is a quark-gluon plasma," he declared. "Period."

In the year since this meeting, a fully fledged feud has erupted between experimentalists and theorists over whether RHIC has indeed created a quark-gluon plasma (often shortened to QGP). The

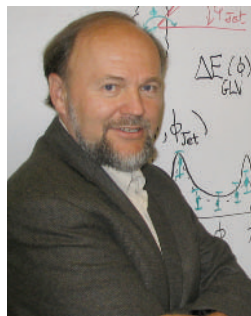
two camps are still working together — the nature of the experiments means that they have to — but they are also locked in an intense and very public debate that both sides agree is unprecedented in the field. The result has been confusion among journalists, government supporters of the work, and interested members of the public.

This summer has brought an uneasy truce. The 1,000-plus physicists and engineers associated with RHIC's detectors reached a collective decision in June to wait for more results before making an announcement. "There was surprising accord," says Mark Baker, acting spokesman for the PHOBOS detector. For their part, the theorists remain unapologetic. The findings pass the "toughest and most critical test," says Gyulassy. "Within our current models, there is no other state that can exist," adds Xin-Nian Wang, who heads the nuclear-theory programme at Lawrence Berkeley National Laboratory in California.

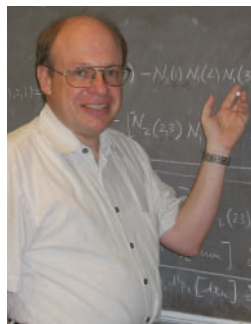
Cosmic goo

The QGP is the name assigned to what experimentalists and theorists alike believe was the earliest form of matter in the Universe. Just a millisecond after the Big Bang, the story goes, the Universe was little more than a hot, gooey mass of quarks — the sub-subatomic particles that make up protons and neutrons, among other things — and gluons, the particles that glue the quarks together. Physicists hope that studying the plasma in the lab will bring them closer to understanding the early Universe.

But to study a QGP, physicists first have to make one. This is no mean feat because in today's Universe, quarks are firmly bound in



Theorists Miklos Gyulassy (above) and Ulrich Heinz believe that a quark-gluon plasma has been created.





Experimentalists at RHIC want more data before naming the new form of matter they have generated.

groups of two or three. To pull them apart you have to smash large atoms together at high speeds, a task that requires an enormous amount of energy. Enter RHIC. It uses several accelerators to push gold atoms up to energies of 100 billion electron volts inside a 4-kilometre ring and then smash them together at six different collision points.

Each head-on collision creates a fireball 300 million times hotter than the surface of the Sun. It's at that moment that physicists believe a QGP might briefly be created. Detectors at four of the collision points record the shower of subatomic particles released from the gold atoms, and computers extrapolate what happened during the impact.

The experimentalists are like accident investigators who have a series of time-lapse photos from a head-on vehicle collision, and must reconstruct what the thousands of working parts inside the two cars looked like the instant their front bumpers touched. Except that, rather than having examples of the actual cars to work with, all they have are sketches from automotive experts of what they think the cars look like.

The biggest problem for the detector teams is that there is no single 'sketch' or model for a QGP, says Baker. At different points in the collision process, different theories must be used to explain how the gold atoms are behaving. Sometimes some of the theories agree with the data, but sometimes they don't. "Until the whole theory everywhere is understood, I'm not ready

to sign off on a QGP discovery," Baker says.

But Ulrich Heinz, a theorist at Ohio State University in Columbus, says that sort of expectation is unrealistic. Converting colliding atoms into a QGP requires a phase transition — like water boiling into steam — and so it is not surprising that different theories are needed to explain the different phases. "We'll always have to live with the fact that

these quark–gluon plasmas are very dynamical beasts," he says.

Further confusion stems from the fact that RHIC has changed the scientists' understanding of the QGP itself. For example, most theorists believed that a QGP would behave like a gas, but RHIC's data suggest that it flows more like a liquid. "It's not what people had predicted," says Nick Samios, a former director of Brookhaven.

The complexity of the experiments lies at the heart of the debate between the two communities. Unlike colliders that smash together individual protons and antiprotons, the gold collisions in RHIC bring some 400 neutrons and protons together at once. Trying to capture what happens as these all fleetingly turn into mush is a challenge both for the detectors and for the subsequent painstaking analysis. Each collision yields billions of data points, which are used to extrapolate a set of properties that only partially resemble a QGP — and whose behaviour is far harder to predict than that of a single particle. Zajc sums up the situation succinctly: "There's no smoking gun."

Such complexities are not confined to

"It's as if Columbus hit land and came back unexcited about it."

— Miklos Gyulassy

RHIC's experiments. In 2000, researchers at the Super Proton Synchrotron at CERN, the European Laboratory for Particle Physics near Geneva in Switzerland, announced that they had tantalizing hints of a QGP. They stopped short of claiming a QGP discovery, choosing instead the term 'quark–gluon matter'. But journalists were less discriminating, and many used the two terms interchangeably. "Many people here felt that they stole the thunder from RHIC," says Baker. RHIC researchers lashed out at CERN, calling the claims unfounded. So, Baker says, the RHIC team feels even more pressure to make careful and conclusive measurements.

Split decision

Another reason why the experimentalists are reluctant to rush into claiming a discovery may be that they won't face any competition until the next accelerator at CERN comes online in 2007, says Samios. Until then there is no pressure to make a big announcement. "It's more psychology than physics," he says.

By contrast, the theorists who have devoted decades to calculating features of the QGP are impatient to declare victory. "To me it's as if Columbus hit land and came back unexcited about it," says Gyulassy. Heinz adds that it is now important to move away from discovery measurements and to begin characterizing the plasma's behaviour. "We need to start exploring the details of this new kind of matter," says Heinz.

What the public would make of another discovery announcement is unclear. Attendees at the annual Quark Matter conference held in Oakland, California, in January were confronted with two seemingly contradictory positions. Experimentalists offered reams of collision data without dubbing the phenomenon a QGP, whereas some theorists, including Gyulassy, announced the discovery of a QGP in their talks. When *The New York Times* ran a story titled "Tests suggest scientists have found Big Bang goo", officials at the Department of Energy, which oversees RHIC, reportedly hit the roof because they had not been consulted before such a major announcement.

Behind the scenes the two camps are still talking, and at a quieter lab meeting in June took the first steps towards a reconciliation. Wang believes that the meeting helped experimentalists focus more on the 'big picture', which may speed a conclusion to the debate.

For now, it seems that both sides are trapped in a stalemate. Zajc hesitates to give a precise date for when, or even if, a QGP discovery will be announced at RHIC. "We're still collecting evidence, and I think we're making good progress," he says. But the theorists say that the days of QGP are already here. "The evidence is overwhelmingly clear," Gyulassy says. "It's time we gave it a name." ■

Geoff Brumfiel is *Nature's* Washington physical sciences correspondent.

When two tribes go to war

Medical geneticists and isolated Native American communities afflicted by inherited diseases should have much to gain from working together. But the relationship can go sour, as Rex Dalton finds out.

South of the Grand Canyon in Arizona, in a valley that roads still don't reach, the Havasupai tribe has for centuries lived a cloistered existence in the high desert. Isolation in a geological wonderland has allowed the tribe's 600-plus current members to protect their ancient culture. But the flipside is a restricted gene pool that has given the Havasupai one of the highest incidences of type 2 diabetes anywhere in the world.

Such populations offer geneticists the chance to discover rare gene variants underlying disease that would be difficult to detect in more diverse groups. And in the early 1990s, with the tribe's blessing, a team from Arizona State University (ASU) in Tempe began searching for a genetic cause of the Havasupai's diabetes.

Instead of a genetic breakthrough, the research project has spawned lawsuits claiming \$75 million in damages, filed by tribal members who claim that their rights were infringed. The accused researchers strenuously deny any wrongdoing, and blame the dispute on a series of misunderstandings. These problems seem to have been inflamed by personal differences among the scientists involved. But the case illustrates the sensitivities associated with conducting genetic research on Native American populations — which, enriched by gambling revenues, are now in a position to assert their legal rights.

Medical geneticists are watching the lawsuits carefully, as some believe the results could cast doubt on the future of genetic studies being conducted on Native American populations across the United States. In the meantime, the Havasupai reservation, some

240 kilometres northwest of Flagstaff, is closed to researchers.

"What concerns me deeply is that the allegations have resulted in a moratorium on biomedical research on the Havasupai reservation, excluding this and other communities from discoveries with the potential to address their health concerns," says Therese Markow, who led the Havasupai project during her years at ASU, and is now at the University of Arizona in Tucson.

Family roots

To investigate the genetics of disease in small, remote populations, it is important to determine just how genetically isolated a group really is. With modern molecular tools, researchers can examine a tribe's genetic history, revealing where tribal ancestors migrated from, and the degree to which they have interbred, over the years, with other groups — including Americans of European extraction.

These are sensitive subjects for Native Americans. Details about migration may challenge the received cultural wisdom about tribal origins, and the question of who is 'more native' can be particularly contentious.

In decades past, Native Americans feeling violated by intrusion into such territory would have had a hard time challenging researchers from a major university. But the balance of power is shifting. Gambling on reservations produces millions of dollars in

revenue, and has made native tribes a political force. Given their remote location, the Havasupai do not run a casino, but they share in revenue from those Arizona tribes that do. And the state's tribes work together on various issues, hiring well-connected lobbyists and high-powered attorneys to protect their collective interests.

The growing influence of Native American tribes has already been used to block the publication of studies deemed culturally offensive — a development that has split researchers working with native communities (see 'The heart of the matter', overleaf). Some see it as unacceptable censorship; others argue that the tribes' cultural sensitivities must come first.

Against this background, the Havasupai experience illustrates just how badly things can go awry. The project began with great promise — melding Markow's interests in genetics with the expertise in social anthropology of John Martin, an ASU colleague who had worked with the tribe for more than 40 years. Martin knew several generations of Havasupai and had created a genealogical history of tribal families.

At the beginning of the twentieth century, a combination of disease and natural disasters had reduced the Havasupai to about 165 members, with only about 40 men and 40 women of reproductive age. As the population recovered from this bottleneck,

"Native American tribes are so understudied. If this litigation continues, all research is going to cease."

— Daniel Garrigan



C. KARNOW/CORBIS



In dispute: the Havasupai tribe initially welcomed geneticists into its remote settlement (right), but subsequent research has led the Native Americans to file lawsuits for \$75 million in damages.



the rate of diabetes began to climb. In 1991, when Markow's study began, 55% of Havasupai women and 38% of the men were diabetic. Martin had also observed a mental condition that he suspected to be schizophrenia — and which his family charts indicated may have originated with a tribal shaman in the late nineteenth century.

On the trail

Markow and Martin put together a proposal to study diabetes, schizophrenia and depression in the Havasupai, which was approved by ASU's human subjects committee in 1991. Initially, the study was conducted with university funds and a grant of some \$90,000 from the National Alliance for Research on Schizophrenia and Depression in New York. Later on, there would be small grants from the National Science Foundation and the National Institutes of Health.

Working with the Havasupai meant hiking, riding horseback or taking helicopters into the reservation. Tribal members who volunteered for the study signed a consent form in which they agreed to provide blood samples, plus hand and fingerprints, for genetic studies into behavioural and medical disorders.

Pursuit of a genetic cause of schizophrenia was dropped early on, after a psychiatrist found no evidence of the disease among tribe members. And as the ASU researchers began to examine Havasupai DNA samples,

they also ran into trouble finding a genetic link to diabetes. Studies of another Arizona tribe, the Pima, had by then found that 81% of nearly 200 Pima diabetics carried a particular variant of a gene involved in immune recognition called *HLA-A2* (ref. 1). But Markow's team could find no association between this gene and diabetes in the Havasupai².

The project looked set to yield nothing but results on the general biological consequences of the tribe's restricted gene pool³ — especially after cell lines created from the Havasupai blood were damaged by a freezer failure in 1994. By the time that Markow moved to the University of Arizona in 1999, it seemed as if all momentum had been lost.

But in 2002, the project was reinvigorated after genetic material was salvaged from the damaged cell lines. Daniel Garrigan, a PhD student in another lab at ASU, was then able to analyse genetic markers called microsatellites — repeating sequences of two or more letters of the genetic code that vary in their length from person to person — to examine genetic variability among the Havasupai.

By early 2003, Garrigan had a manuscript accepted by the journal *Genetics* detailing markers in the Havasupai that were sufficiently variable to use in the search for genes predisposing to diabetes. This was also a major part of his doctoral thesis — until he walked to the podium at ASU on 4 March 2003 to deliver a lecture as part of the examination process.

The audience included Martin and an acquaintance of his from the Havasupai,

Carletta Tilousi. She publicly questioned Garrigan's authority to perform the study. "It was a bizarre event," recalls Garrigan. The day before the lecture, Garrigan says Martin had warned him that he viewed the Havasupai data as his intellectual property. "Stop, or there will be repercussions," Garrigan claims he was told. Martin agrees that he challenged Garrigan on his use of the samples, but denies that this was a threat.

Culture clash

While he had made been aware of Martin's views on the ownership of the data, Garrigan was stunned by Tilousi's intervention at his lecture. Having worked among the Maya in Mexico before joining the Havasupai project, he thought that he was sensitive to Native American concerns. He also believed that his studies had put him on the trail of a genetic explanation for diabetes in the Havasupai.

When Martin and Tilousi objected to Garrigan's research, meetings were held at ASU. A decision was made to withdraw the manuscript from *Genetics*, and to remove the

microsatellite data from Garrigan's thesis. "It was disheartening," says Garrigan, who is now a postdoc at the University of Arizona.

After the thesis showdown, the conflict intensified. Martin first complained to ASU's human subjects committee, alleging that the research had strayed away from the topic of diabetes into areas that the tribe had not agreed to. When the committee investigated this and found no problem, Martin wrote to ASU president



Therese Markow sought a genetic cause for diabetes in the Havasupai.

The heart of the matter

When it comes to genetic studies of Native Americans, anthropologist Robert Williams of Arizona State University in Tempe has learned that — in addition to passing scientific peer review — it may be necessary to pass a test of cultural sensitivity.

After eight years of laboratory work on blood samples taken from 5,000 individuals belonging to 12 tribes across the United States, Williams was in 1999 required to halt his attempt to investigate the growing incidence of cardiovascular disease among the tribes.

His project was part of the Strong Heart Study, a large epidemiological study that has received \$46 million over 15 years from the National Heart, Lung, and Blood Institute (NHLBI) in Bethesda,

Maryland. Williams suspected that the epidemic of heart disease among Native Americans was in part caused by gene flow from interbreeding with Americans of European origin. So he set about investigating this genetic 'admixture', an approach he had previously used in studies of diabetes among Arizona's Pima tribe⁴.

Every manuscript arising from the Strong Heart Study is submitted to tribal representatives for review. Williams' paper, dealing with the sensitive issue of interbreeding with non-Native Americans, was the first to be blocked. "I feel this is political censorship," Williams says. "This has never happened in my 30-year career."

Everett Rhoades, a Native American medical researcher at the University of Oklahoma who

was the chair of the ethics committee that recommended halting Williams' research, declined to comment for this article. But epidemiologist Barbara Howard of Georgetown University in Washington DC, who heads the Strong Heart Study, rejects Williams' complaints of censorship. "I know he is frustrated," says Howard. "He is a good researcher. But the concerns of Native American communities take precedence."

Peter Savage, the NHLBI's director of epidemiology, adds that Native American volunteers can withdraw from research and ask that their samples be destroyed at any time and for any reason. "This is the cost of doing research in special communities," he says.

Michael Crow and other university officials on 11 May 2003. In this letter, he alleged that Markow had misused tribal DNA samples by sending them for analysis of their *HLA* genes to two labs in California.

Martin had by then come to believe that the Havasupai's diabetes was caused by nutrition during fetal development, and his letter claimed: "no genetic research on diabetes genes was undertaken" by anyone. He now acknowledges that this statement was an error.

Regardless of its factual accuracy, Martin's letter came at an inopportune time for ASU. Newly installed as president, Crow was leading an ambitious plan to accelerate research at the university, including genetic studies of Native American tribes. A new state-funded research facility, the Translational Genomics Research Institute in Phoenix, was also developing collaborations with ASU researchers and Arizona tribes.

At the same time, the Havasupai were preparing to go public with a press conference, lambasting ASU over the research project. To head off an embarrassing public row, ASU opted to pay for an investigation into the project by an independent attorney.

In retrospect, that move seems to have been a monumental blunder on the part of the university — one that paved the way for the lawsuits now filed against ASU, the University of Arizona, the two institutions that performed the *HLA* gene analysis, and individual scientists including Martin and Markow.

The investigator was Stephen Hart, a lawyer in Phoenix who has represented tribal governments and formerly served as director of the Arizona Department of Gaming, overseeing Native American gambling operations. His report ran to nine volumes, and was delivered to ASU and the Havasupai late last year.

Hart's report is a compendium of interviews. It contains no firm findings of misconduct, but states that there are "issues"



Native Americans have become increasingly politically active — here they protest outside the football ground of the Washington Redskins.

with how the project was administered, the keeping of records, and whether the tribe realized the full extent of the research that would be undertaken. In March this year, a 150-page summary of the report was submitted to the state court in Flagstaff, in support of two lawsuits claiming that the Havasupai's civil rights were violated when their blood samples were used in the research.

Deeply divided

The lead plaintiff on one of the lawsuits is Tilousi, who was elected last December to the tribe's governing council — the plaintiff on the other suit — as concerns grew about the research. Tilousi says she feels mentally "raped" by the project.

The accused researchers are disturbed by Hart's report. In particular, Markow argues that Hart gave too much credence to the testimony of Christopher Armstrong, formerly a researcher at ASU. Armstrong has since been convicted for dealing cocaine and was being treated for drug and alcohol abuse at the time he was interviewed by Hart.

Armstrong accused Markow of various improprieties, including hiding studies on schizophrenia from the Havasupai. Markow

denies the allegations. Far from exploiting the tribe, she argues that her group did extensive screening for diabetes among tribal members and provided health education.

Martin, whose complaints helped to trigger Hart's investigation, is similarly concerned about the report. Indeed, Martin's protests against Garrigan seem to have rebounded badly on him. His relationship with the Havasupai has been torn apart and the suspension of research on the tribe's reservation has halted Martin's work. "I'm bitterly disappointed," he says. ASU officials are now in the process of collecting any remaining stored blood and DNA so that they can be returned to the Havasupai.

Bartha Knoppers, an expert on informed consent in genetic studies at the University of Montreal in Canada, notes that standards have evolved since the Havasupai project began. But she feels that it is difficult to judge the conduct of the research by today's more explicit standards.

In some other cases, researchers have gone back with new consent forms when extending a project, but this is unusual. In the Havasupai case, ASU's human subjects committee accepted that Garrigan's research was an extension of the previously approved project.

Whatever the outcome of the lawsuits, the researchers at the centre of the storm fear that the publicity generated will threaten future genetic studies of Native American tribes. "They are so under-studied," says Garrigan. "If this litigation continues, all research is going to cease."

Rex Dalton is Nature's US West Coast correspondent.

1. Williams, R. C. *et al.* *Diabetologia* **21**, 460–463 (1981).
2. Markow, T. A. *et al.* *Am. J. Hum. Genet.* **53**, 943–952 (1993).
3. Markow, T. A. & Martin, J. F. *Ann. Hum. Biol.* **20**, 389–394 (1993).
4. Williams, R. C., Long, J. C., Hanson, R. L., Sievers, M. L. & Knowler, W. C. *Am. J. Hum. Genet.* **66**, 527–538 (2000).

For more news and analysis go to
news @ nature.com
www.nature.com/news

If you can lose a driving licence, why not a PhD?

A doctorate is seen as a licence to do science. It should be revocable for misconduct.

Sir—Your Editorial “PhD — club or history?” (*Nature* 429, 789; 2004), about the withdrawal of Jan Hendrik Schön’s doctorate, raises an important point. You argue that a PhD is a “piece of history” and that by revoking Schön’s, as a sanction against fabricating results, the University of Konstanz is treating a PhD as a mere “club membership”.

However, a PhD is seen by the wider scientific community as a de facto ‘licence’ to perform science. It is true that some notable scientists, including the late John

Maynard Smith (see *Nature* 429, 258–59; 2004), have not needed the benefit of a PhD to carry out excellent science. But holders of the degree are perceived by the general public as being experts in their fields. I would argue that knowingly publishing fraudulent results does far more to diminish the status of a PhD than selling the bogus doctorates widely available on the Internet “for only US\$35 plus postage”.

Although Schön’s PhD thesis was found to be free of malpractice, no one would argue that someone passing a driving test

flawlessly could not, at some later date, have their licence revoked should their conduct behind the wheel suggest they are no longer a safe and responsible driver. Given the trust that the scientific and non-scientific communities place in a PhD, it is not unreasonable to withdraw one under certain circumstances — rare though I hope these would be.

Adam G. Hart

Department of Animal and Plant Sciences,
University of Sheffield, Western Bank,
Sheffield S10 2TN, UK

Species problem solved 100 years ago

Sir—In his Turning Points essay “Learning from the Altmeister”, Axel Meyer highlights the 100th birthday this year of the great evolutionist Ernst Mayr (*Nature* 428, 897; 2004). Yet the other major centenary for evolutionary biology has been overlooked, both by Meyer and in your anniversary Commentary article “1904 and all that” (*Nature* 426, 761–764; 2003). Edward Bagnall Poulton’s paper “What is a Species?” (*Proc. Entomol. Soc. Lond.* 1903, lxxvii–cxvi; 1904) was the first to grapple exclusively with the problem of species in an evolutionary framework. Poulton’s paper, a version of his January 1904 presidential address to the Entomological Society of London, laid out the research programme for speciation largely adopted today. (See <http://abacus.gene.ucl.ac.uk/jim/Mim/poulton.html>.)

I do not wish to belittle the work of Mayr and the geneticist Theodosius Dobzhansky — but our impression that they solved the species problem is illusory. They were merely the ones who translated it from the technical literature, enunciating much more clearly than before what had become the prevailing view of species among those who had thought about the problem.

Both Mayr and Dobzhansky were strongly influenced by Poulton, as well as by Poulton’s friend Karl Jordan, also a Fellow of the Entomological Society, and both cited their work. Mayr was to form his views on species and speciation, and particularly the ‘biological species concept’, nearly 40 years after Poulton’s pioneering argument that species were reproductively isolated populations.

Yet the historical links go back even further. In December 1903, shortly before Poulton’s lecture, Alfred Russel Wallace gave him a signed book on mimicry and

speciation. This contained a reprint of Wallace’s 1865 paper on Asian *Papilio* butterflies: the first to recognize that the female mimics of poisonous swallowtails were members of the same species as non-mimetic males.

Wallace also made the first careful analysis by a darwinist of ‘varieties’ below the species level, in particular distinguishing geographic races from reproductively isolated species. Darwin had never provided such an analysis, leaving his definition of species vague. Both Poulton and Jordan worked on *Papilio* butterflies, and cited Wallace’s paper.

So, as well as Mayr’s 100th birthday, we should celebrate the centenary of Poulton’s paper and his gift from Wallace. These events were as epochal in their way for evolutionary biology as was the understanding of the structure of DNA for genetics.

James Mallet

Galton Laboratory, University College London,
4 Stephenson Way, London NW1 2HE, UK

Tight budget should fund benefits, not more posts

Sir—Your News story “Young biologists rejected as NIH budget squeezes training grants” (*Nature* 428, 879; 2004) laments the loss of training positions caused by the stagnant training budget at the US National Institutes of Health (NIH).

The desire to hire all deserving candidates is understandable. However, the scientific community must realize that it is in everyone’s interest to make sure trainees receive adequate compensation, even if that means fewer positions are funded overall.

The Federation of American Societies for Experimental Biology is concerned that funding restrictions will limit the growth

of stipends and benefits for all NIH-supported trainees and fellows.

The NIH Kirschstein–NRSA (National Research Service Award) institutional training grants and individual fellowships are the gold standard for postdoctoral research training in the United States. They identify our most talented and promising young researchers. Despite the prestige of this award, postdoctoral NRSA fellows are not guaranteed access to basic employment benefits. With the median age of postdoctoral fellows rising, these benefits are particularly critical to those who must support families with relatively low salaries and inadequate health and retirement benefits.

The Kirschstein–NRSA programme should put aside extra funds to cover the costs of health and other benefits — even if this requires funding fewer slots. This will enhance the stature of the NIH Kirschstein–NRSA programme and strengthen its ability to recruit exceptional research talent.

This much-needed increase in NRSA benefits will send a positive signal to the best and brightest young scientific minds in the United States. We also favour an increase in the NIH training budget, but given the severe budget constraints, it may be necessary to reduce the number of NRSA postdoctoral positions. There is no doubt that this will be a painful decision. However, we cannot afford to compromise the programme by penalizing these outstanding young scientists.

Robert D. Wells

Federation of American Societies for Experimental Biology, 9650 Rockville Pike, Bethesda, Maryland 20814, USA

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

The decline of China's environment

The spread of agriculture led to deforestation and the growth of towns.

The Retreat of the Elephants: An Environmental History of China

by Mark Elvin

Yale University Press: 2004. 547 pp.

\$39.95, £25

Crispin Tickell

Pity the poor elephants! Over more than 4,000 years they were gradually forced from living all over China to a few protected enclaves near the border with Burma. The main reason was the destruction of their habitat as humans cut down forests and introduced agriculture. Farmers found the dwindling elephant herds a nuisance, as crops were trampled and plundered. Others came to value elephants for military, transport and ceremonial purposes: their ivory was prized and their trunks became a gourmet delicacy. Elephant numbers shrank until they were little more than a memory for most Chinese. Mark Elvin uses the decline of the elephant as an allegory to illustrate the transformation of the Chinese environment to the end of pre-industrial times. Some of the same story can be seen in Africa today.

Elvin's book is not so much an environmental history of China as a collection of its fragments. With copious quotations from Chinese written sources of all kinds, he shows what happened in different places and why. Even if we can see from archaeology that comparable events took place elsewhere, only in China are there such written records, giving a unique account of how it felt to live through them. It was not always a pleasant or edifying process, and as usual the voices of those worst affected will never be heard.

In broad terms, the transformation of the Chinese environment, which was faster in some areas than others, had certain characteristics. First, deforestation made way for agriculture. There was then a bonanza as resources were exploited, species were lost and human numbers rose. This triggered the growth of towns, cities and states with social stratification, followed by increasing competition between them, with war as the spur and the environment sometimes used as a weapon. Better technology was mitigated by mismanagement of resources. Entrapment in limited local circumstances left people vulnerable to change. Finally, there was a greater risk of social and economic collapse affecting society as a whole. Elvin shows the differences clearly in three areas: Jiaying to the south of the Yangzi river; Guizhou in the south, where the Han people gradually displaced the indigenous Miao; and Zunhua in the mountainous northeast.



Stacking the odds: Chinese farmers shaped the environment to suit them by building rice terraces.

Everywhere, control of water was essential. 'Hydraulic despotism' may tell only part of the story, but communities and even states grew partly out of the need to manage this precious and sometimes capricious resource. The struggle to run irrigation systems, limit marine incursions, maintain banks and walls, undertake dredging, cope with floods and storms, and adapt to ever-changing weather patterns is as difficult today as it ever was. With huge populations dependent on particular systems, any change can become increasingly difficult to cope with.

The complexity of Chinese attempts to manage human effects on the environment is remarkable. Even more special are the Chinese beliefs and attitudes towards the environment that have existed over the millennia. Generalizations are bound to be misleading but, in general terms, the Chinese were driven, in Elvin's words, by a desire for rational mastery of the world. They had little hesitation in uprooting forests, redirecting and polluting rivers, destroying natural landscapes and giving political and military needs absolute priority. They had remarkable powers of organization, and ran projects far beyond European capacities at the time. But in doing so, the Chinese paid scant regard to the environment and unwittingly created many long-term problems.

On the other hand, the Chinese had a particularly sensitive respect for nature and natural beauty in all its forms. Even as forests were destroyed, individual trees were singled out for admiration. Heaven and Earth were closely linked, and the line between the natural and the supernatural was blurred. There was a confluence of matter leading

to energy, and energy leading to life, each a product of Bright Force and Dark Force. Dragons and spirits were sometimes seen above the surface in thunder and lightning, and sometimes below it in earthquakes. They formed part of a living world that sustained and punished humans. They even related the behaviour of the weather to human activity, so there was morality in meteorology.

In such a world, it was crucial to divine what the invisible forces felt or did. This could involve sacrificing animals or humans, or burning cracks in the shoulder blades of mammals or the undershells of turtles. In Shang times, such practices had political significance as the ruler was the intermediary between the visible and the invisible world. This was also true in other epochs when the apparatus of authority was given almost divine attributes.

It is as difficult for us to enter into this mental cosmology as into that of our own ancestors in pre-scientific times. Elvin shows that searching for observable and verifiable facts about the world, and putting them to use in programmes of thought, was almost entirely alien to the Chinese. As a result, the shock of change was more abrupt in China than it was in Europe, where the scientific revolution began earlier. Traces of the old thinking may have survived Mao Zedong and persist in fundamental ways today.

The Retreat of the Elephants is not an easy book to read. Some of the quotations seem scarcely relevant, and the whole text could have been usefully pruned. At the end there is an unilluminating venture into equations, as if sustainability could be reduced to an algorithm. Yet taken as a whole, the book is a fascinating, scholarly miscellany of stories,

poetry and ideas from the history of the longest continuous civilization on Earth. The relationship of that civilization with its fragile and often tortured surroundings contains lessons for others — particularly at a time when industrial society in China, as elsewhere, is pressing harder than ever on the environment. This will be a source book, elephants and all, for generations to come. ■

Crispin Tickell is a member of the China Council for International Cooperation on Environment and Development.

The body-plan explosion

On the Origin of Phyla

by James W. Valentine

University of Chicago Press: 2004. 614 pp.
\$55, £38.50

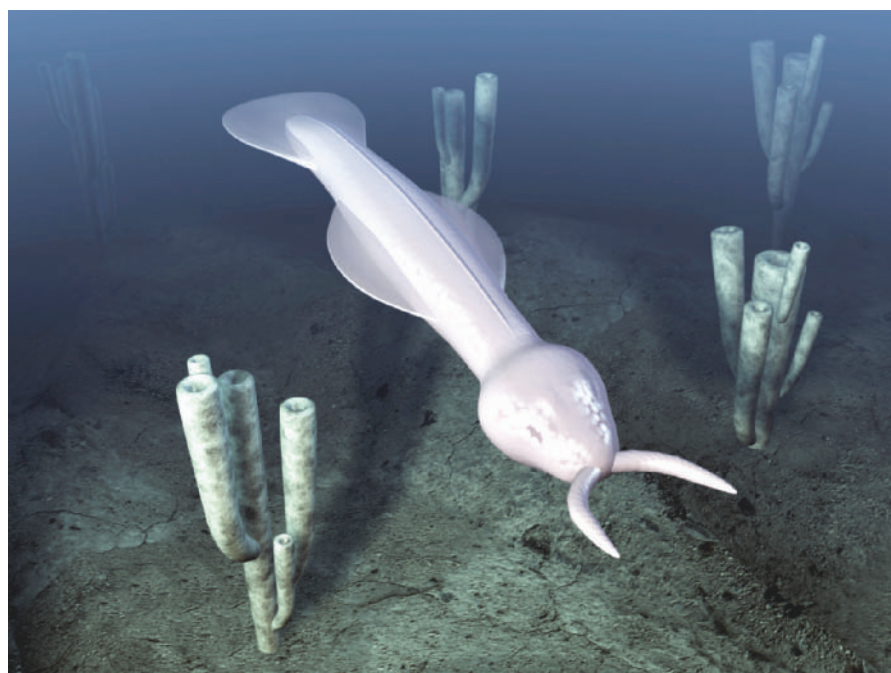
Stefan Bengtson

More than 30 years ago, palaeontologist James W. Valentine's highly acclaimed book *Evolutionary Paleocology of the Marine Biosphere* sought to integrate ecological and environmental studies within a framework of evolution over geological time. He has now written a book on one of the most significant revolutions in the history of life, the Cambrian explosion.

The Cambrian explosion is the transition half-a-billion years ago from a fundamentally microbial biosphere to one in which various multicellular creatures expanded the food web to dimensions and shapes never before realized. It saw the appearance of just about all the major groups of animals that we today recognize as (fossilizable) phyla. Yet Valentine's new book of more than 600 pages contains little about ecology and almost nothing about the physical environment. What has changed since 1973?

For one thing, Valentine, one of the renaissance minds of our times, has further expanded his vision to take in new fields of science, notably developmental genetics, molecular systematics and the evolutionary development (evo-devo) melting pot. Even so, the decision to downplay environmental aspects of early animal evolution must be understood as a deliberate one.

Darwin wisely called his best-known work *On The Origin of Species*; the origin of phyla is an even stickier problem, and Valentine deserves credit for tackling it at such breadth. The first problem is defining the concept of the phylum in an evolutionary context. Historically, phyla have been recognized as groups of extant animals with a characteristic body plan and uncertain relationships to other such groups. Fossils have generally been assigned to the least dissimilar of the living phyla, a practice that pre-empts



In a phylum of its own? Fossils of *Amiskwia* suggest that it was unlike any other known animal.

the possibility of extinct phyla. From a cladistic point of view, it is easy to argue that two extant sister phyla arose at the point of branching from the last common ancestor, and that anything that branched off later should be included in the respective phylum. A distinction can then be made between the crown group (the descendants of the last common ancestor of all the living members of the phylum, including this ancestor) and the stem group (everything else). This avoids the question of how body plans arise and whether there may be others not represented by living forms.

Defining a body plan isn't easy, however. Valentine's definition, for example, is dangerously circular: "an assemblage of morphological features shared among members of a phylum-level group". What does that mean, except that when we define a phylum we also define its body plan, or vice versa? Valentine proposes to define the origin of a phylum by the acquisition of a key apomorphy — a unique derived trait. This may be more subjective and less convenient than letting the total (stem and crown) group or the crown group define the phylum, but it gives due priority to biological significance over methodological convenience. After all, we want to know how different kinds of organism evolve by natural selection, and how they interact with each other and with the environment. They do that with their phenotypes, not their pedigrees.

A key question, then, is whether the body plans of the recognized phyla represent more or less the total number of possible solutions to the problem of being an animal, or whether there were numerous other possibilities that came into being but became extinct because of bad luck or bad design.

Valentine argues that the Cambrian explosion initially produced great disparity in design, but that this was subsequently diminished by extinctions. The pattern of diminishing evolutionary novelty subsequent to this event, he says, may have been due less to developmental constraints than to a saturation effect (candidates for new adaptive radiations were already available among existing body plans). He also believes that the Cambrian explosion produced a lot more homoplasies (similar characters with independent origins) than most phylogenetic analyses suggest — in my view an extremely important point that calls for much more careful character evaluation than is commonly done. He is clearly not impressed, then, by some recent attempts to use fossils to bridge gaps between phyla.

Valentine seems most happy with intrinsic biological mechanisms for the rapid appearance of phyla. Large parts of the book deal with developmental prerequisites (such as cell-type numbers and gene regulation) for the event. Ecological interactions, such as predation, are given more cursory treatment. As for the physical environment, he merely concludes, somewhat apologetically, that although physical environmental factors were "supremely important", he does not see any evidence that extraordinary environmental events were causally connected with the Cambrian explosion. Given that extraordinary environmental events did indeed occur shortly before the explosion, I would give the jury just a little more time to ponder the question. But first I would make sure they had read this magnificent book. ■

Stefan Bengtson is in the Department of Palaeozoology, Swedish Museum of Natural History, Box 50007, SE-10405 Stockholm, Sweden.

Sorrows of the young statistician

Karl Pearson: The Scientific Life in a Statistical Age

by Theodore M. Porter
Princeton University Press: 2004. 352 pp.
\$35, £22.95

Peter J. Bowler

Karl Pearson was one of the founders of modern statistics and a major contributor to the creation of neodarwinian evolutionary theory. He was also a strong supporter of eugenics — the policy of ‘improving’ humans by controlling their reproduction. His contributions to science and mathematics were obscured by the hostility of some of the next generation of statisticians, notably R. A. Fisher. But more recently there has been a wider recognition of the part Pearson played in the rise of a statistical approach that has transformed our vision of nature.

Theodore Porter’s book on Pearson is not a biography in the conventional sense. It focuses on the early part of his career in an effort to show how he was drawn to the study of statistics and eventually conceived it as the key to a new philosophy of nature, which was to become his life’s work. Those seeking extended analysis of Pearson’s work on statistics, evolutionary theory and eugenics will have to look elsewhere, because these have more to do with the later part of his career, which Porter surveys in a single, penultimate chapter. It is not Porter’s intention to carry on, but this could easily



Karl Pearson’s statistical analysis provided early support for Darwin’s theory of natural selection.

be the first part of a two-volume biography.

The great value of Porter’s book is that this focus on the early phase of Pearson’s career highlights the complex route by which his quest for emotional and intellectual satisfaction led him towards the project that would, in effect, create modern statistics. This is very much the story of a romantic hero grappling with the challenges thrown at him by life and love. He studied mathematics at King’s College, Cambridge, emerging as third wrangler in the mathematics Tripos (he was placed third in the lists of those com-

pleting the examination), despite spending much of his time on other intellectual pursuits. These alternative interests led Pearson to study German culture, and he became an expert on the Reformation and passion plays. In 1884, though, he was appointed professor of applied mathematics at University College London, where he worked on the mathematical description of the behaviour of elastic bodies.

This hardly sounds the stuff of high romance, but Pearson was driven by inner passions to seek out a new vision of the truth and make his name by promoting it. In the course of his intellectual wanderings he became a socialist (although the story that he changed his name from Carl to Karl in honour of Karl Marx is apocryphal) and he took up ‘the woman question’ about women’s rights. His relations with the opposite sex were mixed up in complicated ways with his intellectual and ethical interests, but his support for women’s rights was genuine.

In addition to his post at University College London, Pearson was appointed to the Gresham chair of geometry in 1890 and began a three-year course of lectures to non-academic students in which his new approach to statistics was formulated. He chose statistics because by using familiar activities such as games of chance he could appeal to students who would not have appreciated a more abstract approach. For Porter, however, Pearson’s 1892 book *The Grammar of Science* is the key statement of what was now emerging as a foundation for science itself. In 1893, the biologist W. F. R. Weldon approached Pearson for help with analysing his data on variation within

Seeing the world

Our senses allow us to see only a small fraction of our surroundings because we can detect only a narrow band of the electromagnetic spectrum. For more than a century we have developed technologies that have allowed us to create images using energies that exist outside these limits.

In *Invisible Worlds: Exploring the Unseen* (Weidenfeld & Nicolson, £20), Piers Bizony has collected together a sample of the pictures produced by techniques such as cloud chambers, electron microscopes and modern telescopes. Along with more familiar scientific images, Bizony has included examples of the application of these technologies to our everyday lives, such as an ultrasound scan of an unborn baby and this snapshot of illegal immigrants, taken from a covert surveillance vehicle using backscatter X-ray technology. The text that accompanies each picture explains the science behind the imaging technique.



Painting by numbers

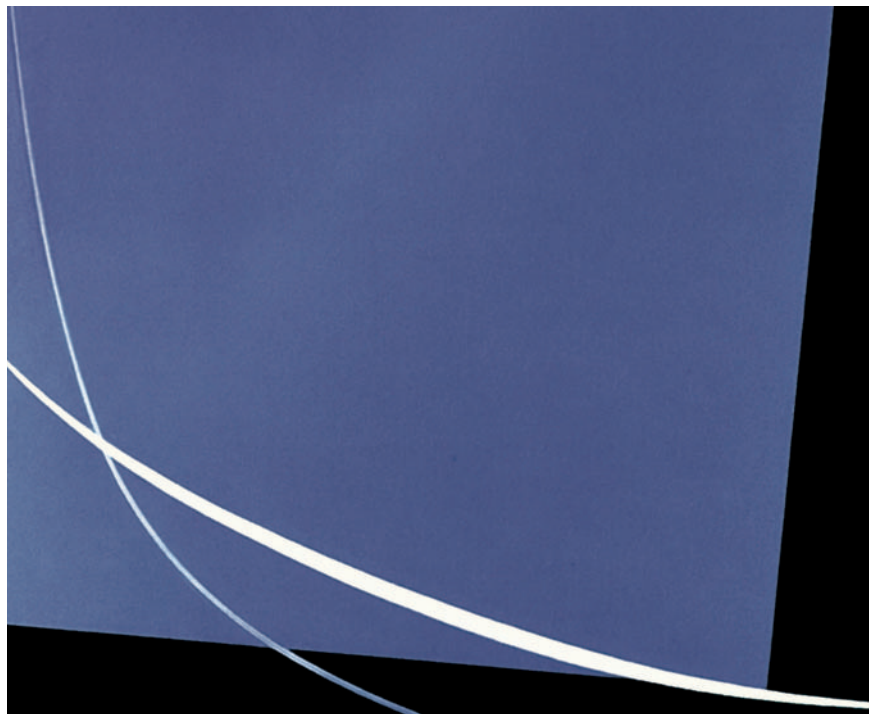
Margaret Leiteritz turns scientific graphs into painted diagrams.

Martin Kemp

Graphs are not generally considered to be things of beauty. Some of the lines that map typical behaviours of physical systems, such as the bell-curve beloved of statisticians, may have a certain visual charm, but the overall appearance of a graph in a scientific publication is not expected to incite aesthetic rapture. However, one of the jobs of artists is to see potential in, and take inspiration from, the most unexpected places, causing us to look afresh at something we take for granted. This is just what the German artist Margaret Leiteritz hoped to achieve with a series of paintings based directly on graphs in texts on chemistry and physics.

Educated first as a librarian, Leiteritz studied between 1928 and 1931 at the Bauhaus in Dessau, the school that was radically reshaping architectural thinking and all related fields of design. Her graduation diploma was signed by such luminaries as Mies van der Rohe, who directed the Bauhaus, Wassily Kandinsky, the Russian pioneer of abstract painting, and Paul Klee, the intellectual master of new and quirky pictorial vocabularies. This gives us an idea of the experimental environment in which she was immersed. After working in Dresden and Wuppertal, she was for many years librarian at the Engler-Bunte Institute at the University of Karlsruhe. But she continued to create art, pursuing a virtually secret career in which she developed a unique kind of painting.

Between 1961 and 1974, Leiteritz produced a remarkable series of 'painted diagrams', derived from a wide variety of the scientific articles and books in her care. She was particularly attracted to chemical engineering, and was fascinated by graphs that recorded phenomena of gases, liquids and solids, such as mixtures, solutions, reactions, liquefactions, viscosity, ignition and combustion. She transformed the linear diagrams into colourful, dynamic and engaging paintings, creating suggestive analogues for the life of the materials during the transformations they were undergoing. In effect, she created portraits of these processes, translating the conventions of graphs into pictorial



forces — much as Klee had translated the energies and motions of the Universe into line and colour, aspiring to create psychologically effective signs in a new kind of pictorial language. Her paintings, like those of Klee, are abstracted, but not abstract in the sense of not representing something.

One of Leiteritz's most elegant graph paintings, *Crossing at the Left Border* (shown here), painted in 1966, was prestigiously used on the cover of the catalogue for the exhibition *50 Years Bauhaus* in Chicago in 1969. It was developed from figure 13 — "Separation as a function of flow rate with plate spacing as parameter" — of a paper entitled "Separation of liquids by thermal diffusion" by J. E. Powers and C. R. Wilke (in *Am. Inst. Chem. Eng. J.* 3; 1957). The mellifluous curves are closely transcribed from her source, and the colours and their tilted borders are designed to convey a sense of the dynamic separation processes induced by temperature. The materials that Leiteritz used —

oil on a linen-lined aluminum plate — sustain the technical tone of the image.

Leiteritz was trying to express that the graphs in the books and periodicals that passed through her hands were not inert and uninvolved records from specialist nooks and crannies of science; they were precious visualizations of the underlying actions of materials in nature, embodying the forces and forms that make up the patterns and rhythms of the world. To use a philosophical distinction, she said: "The diagram is not the thing in itself, but represents something that happens."

Her pictures can be seen at the Museum for Applied Art in Frankfurt, Germany, until 5 September, and are illustrated in *Gemalte Diagramme: Die Bauhaus Künstlerin Margaret Leiteritz* (Inno Verlag, 1993).

Martin Kemp is professor of the history of art at the University of Oxford, Oxford OX1 1PT, UK, and co-director of Wallace Kemp Artakt.

populations, as a test of the darwinian theory of natural selection. The pieces of the jigsaw puzzle now fell into place, and Pearson was soon pioneering his revolution in statistical thinking.

The end results of Pearson's intellectual pilgrimage are clear enough, but specialists may have doubts about the way Porter sees the whole package eventually coming together. Porter is not over-generous with references to the secondary literature, which makes it difficult to relate his interpretations to those published by other scholars. It is not always clear just how revolutionary Pearson's

new techniques were: Porter writes of him taking up mathematical statistics as though the field were already in existence, but some would argue that Pearson actually created the field in its modern form. Furthermore, Eileen Magnello has suggested (*Hist. Sci.* 37, 79–196, 123–150; 1999) that *The Grammar of Science* is not the best guide to the whole statistical project. And Weldon plays a disappointingly small role in Porter's book, considering that it was the problems generated by his data that helped Pearson complete his intellectual odyssey.

Porter's study is innovative because its

focus on the complex nature of Pearson's early life provides a major counterweight to conventional studies of his later work in biology and eugenics. Whether his intellectual and emotional struggles made him the tortured outsider that Porter envisages is a moot point — most turn-of-the-century intellectuals were that way inclined, at least in Britain. But perhaps we should accept that with so complex a figure, differing interpretations are bound to emerge. ■

Peter J. Bowler is professor of the history of science in the School of Anthropological Studies, Queen's University, Belfast BT7 1NN, UK.

Why can't planets be like stars?

Planetary science: both the deductive skills of geologists and the mathematical approach of astrophysicists are needed to study planets.

Stuart Ross Taylor

Planets are diverse individuals formed by stochastic processes. In our Solar System we have eight planets, all of which are distinct from one another in mass, density, composition, rotation rates and angle of inclination (obliquity). Their only common properties are near-circular orbits and low inclinations to the Earth–Sun plane, characteristics that enabled Pierre-Simon Laplace to conclude in 1796 that they had originated from a rotating disk, the solar nebula.

Our planetary system also includes more than 120 moons and a host of smaller bodies, most of which exhibit some peculiarities of composition or behaviour. Thus, there appears to be no uniformity in the processes of planetary or satellite formation from the gases, ices and rocky components of the primordial nebula. Planets may also migrate from their original positions, effectively randomizing any initial radial variations in the chemistry of the nebula.

The discovery of more than 100 planets orbiting stars other than the Sun has brought the question of planetary origin and evolution into sharp focus. Our limited sampling of these extrasolar planets reveals even wider variations in terms of mass and spacing of planets and — to add additional complexity — many of these newly discovered planets are in highly elliptical orbits. It seems likely that we may eventually find that planets forming from disks rotating around young stars will occupy all available niches within the limits imposed by the cosmochemical abundances of the elements and the laws of physics and chemistry.

Unlike planets, stars are relatively uniform in composition and differ mostly in mass. The basic features of stellar evolution have long been understood and are described by the Hertzsprung–Russell diagram, which relates surface temperature to luminosity and essentially reflects the nuclear reactions taking place in stars.

By contrast, planets are individuals that show few systematic relationships and have resisted attempts at classification or even definition, as witnessed in the furor over the status of Pluto, which is an eccentric dwarf when placed among the planets, but is better suited to be the king of the many icy bodies in



Moon talk: images of Phoebe shed light on the formation of planets.

the Kuiper belt. So far, there is no planetary equivalent of the Hertzsprung–Russell diagram. Even if we arrive at a satisfactory explanation for the formation and evolution of our planetary system, there is no guarantee that this will apply elsewhere. Perhaps this is the reason, as Stephen Brush has commented, that the origin of the Solar System represents one of the oldest unsolved problems in science.

The problems of studying planets are well illustrated by the history of attempts to understand the Earth. Geology was a late-comer among the sciences. Even after James Hutton's insights into deep time in 1788, it was a further 150 years before plate tectonics was understood as the mechanism responsible for the architecture of the Earth's surface. Plate tectonics has the useful property of both building continents and forming ore deposits useful for advanced civilizations — in doing so, enabling this discussion to take place. But this process is unique to the Earth among the planets of the Solar System and was only made possible by the late stochastic addition of a water content of a few hundred parts per million. Many of the difficulties in trying to understand the evolution of the Moon arose from the uncritical attempts to apply our hard-won experience with wetter terrestrial rocks to those from our bone-dry satellite.

Even when nature got around to building two similar planets, it finished up with the Earth and Venus. These twins are close in mass, density, bulk composition and the abundances of the heat-producing elements (potassium, uranium and thorium).

But Venus is a one-plate planet without a moon and seems to undergo planetary-wide resurfacing with basalt perhaps once every billion years. What causes the difference between the geological histories of these twins? The short answer is water, but much may be due to variations in the early history of impacts during planetary accretion. Similarity is not identity, and the Earth resembles Venus much as Dr Jekyll resembled Mr Hyde. As we search for terrestrial-like planets elsewhere, we need to find out the reasons for these differences and the conditions that allow these diverse bodies to form at all.

The study of planets represents a new area in scientific enquiry, just as geology did in the nineteenth century. While one might intuitively think that the rocky planets could be left to the geologists and the gas giants to the astronomers, neither group alone seems well suited to producing the necessary synthesis. Geologists deal mostly with surface observations, but planetary crusts differ greatly from interiors. In contrast, astronomers have long been involved with the internal evolution of stars. No single specialist group seems well equipped to handle the diversity of planetary systems and the philosophical problems in dealing with stars, the Earth and the variety of solar and extrasolar planets. A unifying approach is needed to avoid the dilemma of the six wise but blind men confronted with an elephant. Just as astrophysics, geochemistry, biochemistry and geophysics have risen at the boundaries between the classical sciences, so planetary science now requires new types of investigators. Such investigators need a distinct mindset somewhere between the approaches of astronomers — who want to treat planets mathematically like stars — and geologists, who want to generalize from their detective-like experience with the Earth. ■

Stuart Ross Taylor is in the Department of Earth and Marine Sciences, Australian National University, Canberra ACT 0200, Australia.

FURTHER READING

Brush, S. G. *A History of Modern Planetary Physics* Vol. 3 (Cambridge Univ. Press, 1996).
Stevenson, D. J. *Phys. Today* **57**, 43–48 (2004).
Taylor, S. R. *Destiny or Chance: Our Solar System and its Place in the Cosmos* (Cambridge Univ. Press, 2000).

Odorant receptors make scents

Rainer W. Friedrich

The goal of making sense of the sense of smell has come a step closer. Work on fruitflies reveals that odorant receptors act as bidirectional chemical detectors and determine the function of sensory neurons.

Olfaction, once thought to be a primitive sense, is now recognized as an elaborate sensory system that deploys a large family of odorant receptors to analyse the chemical environment. Interactions between these receptors and their diverse ligands translate the world of odours into a neural code, but the mechanisms governing this complex process are not totally understood. More than a decade after the discovery of odorant-receptor genes in rodents¹, a study by Hallem, Ho and Carlson², reported in *Cell*, now provides a missing link between the molecular biology of odorant receptors and the physiological properties of sensory neurons.

Odorant-receptor proteins are exposed to the outside world on the membranes of sensory neurons. Interaction of a receptor with an odorant ligand results in a biochemical signal inside the cell that ultimately controls the generation of electrical impulses. These action potentials are transmitted down the axon of the sensory neuron to the brain, where they are integrated with other inputs. However, the mechanisms that tailor this transduction process to produce responses that are specific to particular odours are not completely understood. For example, is odorant specificity directed solely by the odorant receptor, or does it depend on other proteins?

Such questions would normally be tackled by performing experiments in which the genes for the receptors are expressed in artificial systems³, but odorant receptors have proven notoriously reluctant to submit to

this approach. However, odorant-receptor genes can be functionally expressed in their natural environment, the olfactory sensory neuron⁴. Hallem and colleagues² therefore played a genetic trick and used a type of olfactory sensory neuron called ab3A as an expression system in the intact fruitfly, *Drosophila*. They deleted the genes for the natural odorant receptors from ab3A, and replaced them with genes encoding other odorant receptors, one gene at a time (Fig. 1). They then measured the action potentials generated by the ab3A neurons in response to a panel of different scents.

Removing the ab3A receptors abolished all odour sensitivity in the neurons⁵, but in many cases expression of a single odorant receptor reinvigorated the neurons. Different types of sensory neuron were previously defined according to their physiological responses, but the identity of the odorant receptors expressed on each neuron type remained a mystery. In the genetically altered ab3A neurons, many receptors produced responses that were essentially indistinguishable from those displayed by one of the previously defined neuron types. The simplest explanation for this striking result is that the functional properties of a sensory neuron are determined by the receptor protein, and that each sensory neuron expresses only one functional odorant receptor, which is consistent with results in rodents⁶. Some of the receptors yielded responses that did not match known response spectra, and thus may be expressed in unknown neuron types. Finally, some odorant receptors produced no

response to any of the tested compounds, and therefore might encode receptors that detect specialized chemical signals with only a very narrow ligand range.

In normal sensory neurons, most responses are excitatory, but some are inhibitory. Hallem *et al.*² found that the ab3A neuron could respond with either excitation or inhibition to a given odorant, depending on which odorant receptor was expressed. Moreover, the same receptor stimulated by different scents could mediate both excitatory and inhibitory responses, and the odorant receptor determined the neuron's spontaneous action-potential firing rate. It thus seems that different response directions are not mediated by different transduction mechanisms, but that odorant receptors can signal in both directions. The authors propose a model in which a receptor exists in equilibrium between an active and an inactive state that is shifted gradually in either direction by ligands. So we will need to refine our thinking, and view odorant receptors as bidirectional chemical detectors.

The chemical information that can be sensed by an organism is determined by the receptive range of its odorant-receptor family. Consider the analogy of colour vision: in humans, three types of photoreceptor span the visible spectrum and allow colour discrimination. Knowing the wavelength spectrum of each photoreceptor type therefore provides insight into how colour is coded into neural signals. By comparison, the olfactory system is much more complex: humans have roughly 350 odorant receptors, and

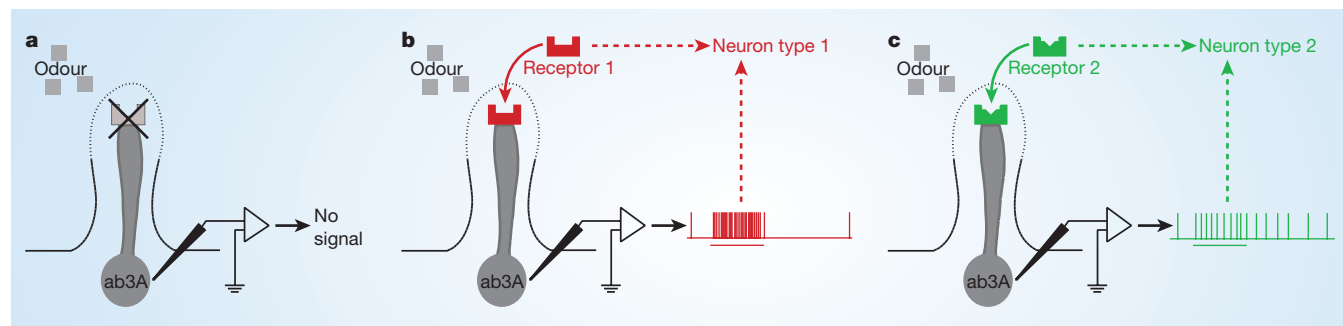


Figure 1 Linking odorant receptors to neuronal responses. Hallem *et al.*² expressed odorant receptors in the olfactory sensory neuron ab3A in bristles of the *Drosophila* antenna. The odorant receptors that are normally found on ab3A were deleted genetically (a), and substituted by receptors from other neurons (b,c). In many cases, expression of the foreign odorant receptors re-established the

neurons' responses to a variety of odours. The intensity and kinetics of the neurons' responses depended on the identity of the receptor and matched those of known sensory neuron types (b,c). These results demonstrate that the odorant receptor determines the specificity of olfactory sensory neurons and establish a receptor-to-neuron map in the *Drosophila* antenna.

rodents have more than a thousand. A complete characterization of the repertoire of olfactory receptive ranges will require measurement of the responses of all receptors to all possible ligands. *Drosophila* seems to be the best model for such a venture because it has only about 60 odorant receptors⁷. These receptors are divided between two distinct chemosensory organs, the antenna and the maxillary palp. In their study², Hallem and co-workers have characterized 31 of the 32 receptors expressed in the antenna. So perhaps the greatest merit of their work is its comprehensiveness: it describes, in remarkable detail, the receptive ranges of almost an entire chemosensory organ, albeit to a (necessarily) limited set of odours.

Most of the odorant receptors tested responded to a relatively broad, but nevertheless specific, spectrum of ligands. This is consistent with the response properties of sensory neurons in other organisms and with the tuning of neurons in the first olfactory processing centre in the brain. Precise odour information is therefore encoded combinatorially in activity patterns across multiple neurons. This notion was formulated

more than 50 years ago⁸ and has since been examined in detail in many organisms⁹. The study by Hallem *et al.*² provides a molecular basis for this view of olfactory coding. Not only is this necessary for understanding the link between odorant receptors and the neural output of sensory neurons, but it will also allow further studies of receptor–ligand interactions — interactions that ultimately constrain olfactory coding strategies and constitute the keyholes through which the brain views the world of odours. ■

Rainer W. Friedrich is at the Max Planck Institute for Medical Research, Department of Biomedical Optics, Jahnstrasse 29, 69120 Heidelberg, Germany. e-mail: rainer.friedrich@mpimf-heidelberg.mpg.de

1. Buck, L. & Axel, R. *Cell* **65**, 175–187 (1991).
2. Hallem, E. A., Ho, M. G. & Carlson, J. R. *Cell* **117**, 965–979 (2004).
3. Krautwurst, D., Yau, K. W. & Reed, R. R. *Cell* **95**, 917–926 (1998).
4. Zhao, H. *et al.* *Science* **279**, 237–242 (1998).
5. Dobritsa, A. A., van der Goes van Naters, W., Warr, C. G., Steinbrecht, R. A. & Carlson, J. R. *Neuron* **37**, 827–841 (2003).
6. Malnic, B., Hirono, J., Sato, T. & Buck, L. B. *Cell* **96**, 713–723 (1999).
7. Vossahl, L. B., Wong, A. M. & Axel, R. *Cell* **102**, 147–159 (2000).
8. Adrian, E. D. *Acta Physiol. Scand.* **29**, 5–14 (1953).
9. Friedrich, R. W. & Stopfer, M. *Curr. Opin. Neurobiol.* **11**, 468–474 (2001).

Superconductivity

Why the temperature is high

Jan Zaanen

According to a new empirical law, the transition temperature to superconductivity is high in copper oxides because their metallic states are as viscous as is permitted by the laws of quantum physics.

Dissipation is obvious in the human environment. The phenomenon describes how useful energy is eventually converted into microscopic disorder, which is perceived by us as a rise in temperature. But viewed from the fundamental perspective of quantum physics, dissipation is not at all obvious. A striking example is the superconductor — a quantum state of matter in which electrical currents flow without friction. Heat is the enemy of this state and above a certain temperature, the transition temperature, dissipation takes over again. Bardeen, Cooper and Schrieffer's 1957 explanation of superconductivity (in terms of paired electrons) seemed to be one of the great triumphs of twentieth-century physics — until the discovery in 1986 of a new class of superconductors with very high transition temperatures¹. Despite years of intense research, these high-temperature copper-oxide superconductors are still on the list of the great mysteries of physics².

On page 539 of this issue, Homes *et al.*³ report their discovery of a simple but counter-intuitive empirical law for superconductors, a law that is so general it applies equally well to conventional and to high-temperature

superconductors. The law (let's call it Homes' law) states that transition temperature is proportional simply to the strength of the superconducting state at zero temperature (the superfluid density) multiplied by the quantity that expresses how efficiently electrical currents are dissipated in the normal state above the transition temperature (the electrical resistivity). The ramifications of this law for the copper-oxide superconductors are interesting. Although their transition temperatures are high, the superfluid densities of these superconductors are much smaller than those of the conventional superconductors. Why are the high-temperature superconductors so successful at fighting heat? Homes' law implies that it is because their normal states are extremely dissipative. In fact, according to the laws of quantum physics, it is impossible for any form of matter to dissipate more than these metals do; their transition temperatures are as high as they can be, given the ineffectual nature of the zero-temperature state.

Homes' law is exactly the kind of thing that physicists like: it is simple, quantitative, general, but at the same time surprising. It is no surprise, though, that transition temperature

is connected to the superfluid density — many copper-oxide superconductors are already known to obey Uemura's law, in which the two quantities are simply proportional^{4,5} (all equations are given in Fig. 1). Instead, Homes' law relates the superfluid density to the product of transition temperature, conductivity (which is the inverse of resistivity) in the normal state at the transition temperature, and a universal constant (which has a value of roughly 40). Homes' law is valid when Uemura's law fails, even for conventional superconductors. The conductivity term reflects the capacity of the normal state to dissipate electrical currents, but why is it this quantity that ties the zero-temperature state (the superfluid density) to the transition temperature? Even for an expert this is puzzling. Although Homes' law can be rationalized for both high-temperature² and conventional^{6,7} superconductors, the kinds of argument needed in each case are utterly different.

Homes' law has a deep but simple meaning in the case of high-temperature superconductivity (in conventional superconductors it is much more complicated). Its subtlety is clear through the straightforward technique of dimensional analysis: both sides of the equation should be expressed in the same units, and these units are inverse seconds squared, or s^{-2} . Starting on the left-hand side of Homes' equation (Fig. 1), what has the strength of the superconductor, its superfluid density, to do with time? Well, electromagnetic radiation cannot enter a superconductor when its frequency is lower than the 'superconducting plasma frequency' (which has units of s^{-1}). The square of this quantity is a quantitative measure of the strength of the superconductor (it can be expressed in terms of the density of electrons participating in the frictionless currents) and has units of s^{-2} .

Turning to the right-hand side of Homes' equation, the normal-state conductivity quantifies dissipative electrical transport. This conductivity can also be related to a plasma frequency, but this time the plasma frequency is associated with the density of mobile electrons in the normal state; another poorly understood empirical relation, Tanner's law⁸, insists that in high-temperature superconductors the density of mobile electrons in the normal state is four times the density in the superconducting state. In the normal state, there is another timescale, as well as the plasma frequency: it takes a characteristic period of time (the relaxation, or inelastic-scattering, time) to dissipate electrical currents into heat. So conductivity has the dimension of inverse seconds, corresponding to the square of the plasma frequency multiplied by the relaxation time (Fig. 1).

To balance Homes' equation dimensionally, we need one more factor on the right-hand side with dimension inverse seconds. This must come from the transition temperature. Temperature is easily converted into

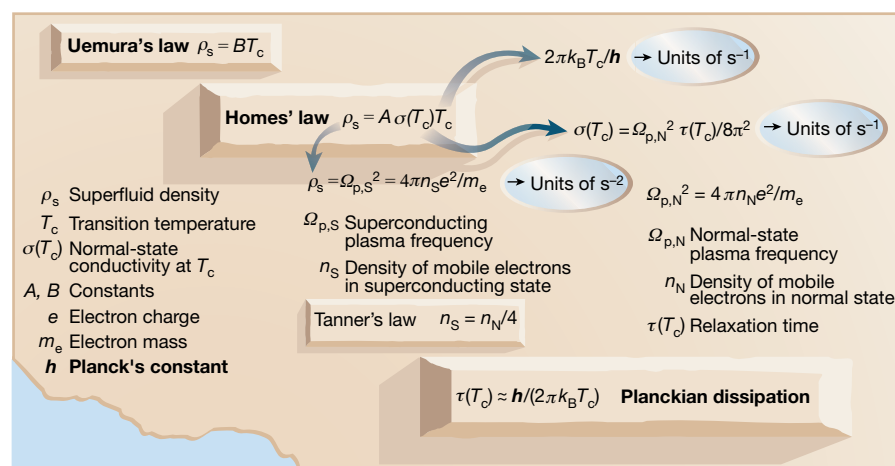


Figure 1 Planckian dissipation and high transition temperatures. Homes *et al.*³ have established a universal relation between the superfluid density, the normal-state conductivity and the transition temperature for high-temperature, copper-oxide superconductors. Further insight comes from a dimensional analysis of the equation: given that both sides of the equation must have units of s^{-2} , the identification of the transition temperature with units of inverse seconds brings Planck's constant into play. This quantum connection is the crux of the matter. Bearing in mind Uemura's law^{4,5}, the quantum physical constraint on the relaxation time — planckian dissipation — explains why the transition temperature for copper oxides is so high.

units of energy, through Boltzmann's constant (k_B). But to convert energy into time requires quantum physics: the uncertainty principle relates energy and time through Planck's constant, h . Putting all these pieces together, we arrive at a rather surprising outcome: Homes' law reduces to the statement that the characteristic timescale for dissipation in the normal state of high-temperature superconductors arises from expressing the transition temperature in units of time through Planck's constant.

This timescale turns out to be very, very short — in fact, the laws of quantum physics forbid the dissipation time to be any shorter at a given temperature than it is in the high-temperature superconductors. If the timescale were shorter, the motions in the superfluid would become purely quantum mechanical, like motion at zero temperature, and energy could not be turned into heat. In analogy with gravity, this timescale could be called the 'Planck scale' of dissipation (or 'planckian dissipation'). That the normal electron fluid in high-temperature superconductors is at the quantum limit of dissipation does not come as a surprise. To reach this limit, the quantum system has to fulfil very specific requirements⁹ — it must be 'quantum critical', with dynamics that seem the same on all scales in time and space. There is, in fact, evidence for the quantum-critical nature of the normal state in high-temperature superconductors¹⁰, including an independent confirmation of planckian dissipation.

What is so surprising about Homes' law is that it relates planckian dissipation to the transition temperature. Uemura's law had already signalled the connection of the superfluid density to the transition temperature, through a simple constant. But what sets the

value of that constant for each compound? Uemura's law and Homes' law are both valid for high-temperature superconductors, so Uemura's constant of proportionality must match the corresponding term in Homes' law — the normal-state conductivity multiplied

by a universal constant (Fig. 1). Because the normal state is a planckian dissipator, with conductivity as small as is permitted by Planck's constant, the transition temperature for copper oxides is consequently high.

Uemura's law, Tanner's law, planckian dissipation — these are extremely simple, empirical relations, particular to high-temperature superconductivity. Conventional superconductivity lacks this kind of simplicity. This is even true for Homes' law. Although it applies to conventional superconductors, it works there for entirely different — and far more complicated — reasons than planckian dissipation^{6,7}. Why should there be this extraordinary simplicity for high-temperature superconductivity? We have as yet no clue, which is why this phenomenon is still on that list of mysteries. ■

Jan Zaanen is at the Instituut Lorentz for Theoretical Physics, Leiden University, PO Box 9504, 2300 RA Leiden, The Netherlands. e-mail: jan@lorentz.leidenuniv.nl

1. Bednorz, J. G. & Müller, K. A. *Z. Phys. B* **64**, 189–193 (1986).
2. Orenstein, J. & Millis, A. J. *Science* **288**, 468–474 (2000).
3. Homes, C. C. *et al.* *Nature* **430**, 539–541 (2004).
4. Uemura, Y. J. *et al.* *Phys. Rev. Lett.* **62**, 2317–2320 (1989).
5. Emery, V. J. & Kivelson, S. A. *Nature* **374**, 434–437 (1995).
6. Tinkham, M. & Ferrel, R. A. *Phys. Rev. Lett.* **2**, 331–333 (1959).
7. Ambegaokar, V. & Baratoff, A. *Phys. Rev. Lett.* **10**, 486–489 (1963).
8. Tanner, D. B. *et al.* *Physica B* **244**, 1–8 (1998).
9. Sachdev, S. *Quantum Phase Transitions* (Cambridge Univ. Press, 1999).
10. van der Marel, D. *et al.* *Nature* **425**, 271–274 (2003).

Cell biology

How to build a cell junction

William I. Weis

Structures of the protein vinculin reveal drastic conformational changes associated with binding to its partners in cell-adhesion contacts. These changes might let vinculin regulate the assembly of these complexes.

The development and organization of tissues depends on adhesive junctions, structures that enable cells to stick to one another and to attach to the extracellular matrix. These highly dynamic multi-protein complexes reach from the cytoplasm through the cell membrane to the outside of the cell. Linking intercellular adhesion proteins to the cytoskeleton enables, for example, remodelling of epithelial sheets during development. Similarly, cell migration depends on coordinating adhesion to the extracellular matrix with changes in actin polymerization. Papers by Liddington and colleagues (Bakolitsa *et al.*, page 583 of this issue¹) and by Izard and co-workers^{2–4} show how vinculin, a protein found in the cytoplasmic regions of certain types of junction, might regulate the assembly of these complexes.

Vinculin is found in two varieties of junction: adherens junctions and focal adhesions. Adherens junctions mediate

cell–cell adhesion using proteins called cadherins that span the cell membrane, contacting cadherins on other cells and linking to the actin cytoskeleton inside the cell. Focal adhesions, meanwhile, fasten cells to the extracellular matrix, with other proteins — integrins — forming a bridge across the cell membrane from matrix to cytoskeleton.

Vinculin consists of a 'head' region (Vh) of about 850 amino acids linked to a 'tail' (Vt) of roughly 200 amino acids. When the head and tail regions are made as separate proteins, they can bind to other components of the junction complexes. Vh binds to talin, a protein that interacts with the tails of integrins in focal adhesions; it can bind to α -catenin, a homologue of vinculin found in adherens junctions; and it can bind to α -actinin, another actin-binding protein found in both focal adhesions and adherens junctions. The Vh–Vt linker is a binding site for VASP, a regulator of actin polymerization. And Vt binds to filamentous actin.

However, in full-length vinculin an intramolecular interaction between Vh and Vt inhibits binding to these partners^{5–7}. In the cell, then, there must be factors that disrupt the Vh–Vt interaction, thereby activating vinculin for binding to its junctional partners. One such factor is an acidic phospholipid known as phosphatidylinositol-4,5-bisphosphate, which binds to Vt and can disrupt the Vh–Vt interaction to enable the binding of vinculin to talin and actin^{8–10}.

The new structures of vinculin^{1–4} reveal a level of structural plasticity that might be expected of an allosteric protein, where the binding of one ligand leads to conformational changes that affect the binding of another. Full-length vinculin contains five α -helical domains^{1,2}. In this inactive structure, each of the first three domains (D1–D3; see Fig. 1 on page 583) comprises two four-helix bundles that share a central long α -helix, whereas domain 4 is a single four-helix bundle that is connected to Vt (domain 5) by the proline-rich linker. D1–D3 form a clamp that grabs Vt, holding it in a conformation that occludes the ligand-binding sites on Vt and the linker. This Vh–Vt interaction partly blocks the phospholipid-binding site, which is consistent with the proposed role of these compounds in activating vinculin¹.

The transition from this closed, inactive conformation to an open, ligand-bound form involves an enormous number of structural changes. The structures of D1 bound to talin peptides show, remarkably, that the first half of D1 changes from a four- to a five-helix bundle on binding, with a simultaneous repacking of the hydrophobic core^{3,4}. This 'helix bundle conversion' produces a conformation that is not compatible with the Vh–Vt interface. So, talin alters the structure of D1 rather than directly competing for the Vt-binding surface. The structural conversion of D1 on binding talin is reflected in slow binding kinetics and increased stability^{1,11}. The binding of ligands to Vt produces changes in protease sensitivity¹², which suggests conformational changes here too. Moreover, the relative positions of the four-helix bundles comprising the second part of D3 and D4 in vinculin differ substantially from those seen in structures of the homologous region of α -catenin, suggesting that there is significant inter-domain flexibility in this part of the structure¹. These structural observations demonstrate that both intra- and inter-domain structural changes are associated with activation and ligand binding.

What provides the energy needed to remodel and activate vinculin? Experiments with Vh and Vt produced as separate proteins led to the notion that binding of a single partner might compete with the intramolecular Vh–Vt interaction. In full-length vinculin, however, Vh and Vt are covalently linked, which increases their affinity for one another to the sub-nanomolar range¹. It is unclear

whether any one of the vinculin partners is present at a high enough concentration to overcome the intramolecular vinculin interaction, because they each bind to Vh or Vt with an affinity at least 10–1,000-fold weaker than the Vh–Vt binding¹. For example, vinculin-binding talin peptides must be added at extremely high concentrations (250–500 μ M) to activate full-length vinculin *in vitro*¹¹. Similarly, it is not known how the concentrations of phospholipids used to activate vinculin *in vitro* relate to the concentrations at an assembling cell junction. Bakolitsa *et al.*¹ argue that the binding energy of several partners is needed to overcome the thermodynamic and perhaps kinetic barriers to activation.

Viewed in this way, vinculin functions as a logical AND gate, in which binding of two partners is required to generate an output, in this case a stable multi-protein complex. Bakolitsa *et al.* suggest that this property is essential for vinculin to sense the correct cellular localization of its partners, because forming a functional complex with, for example, talin and actin in focal adhesions requires these proteins to be present in the same location and at concentrations sufficient to shift the conformational equilibrium to the open state. The structural changes that occur on ligand binding might create kinetic barriers to reassociation of the head and tail, which would facilitate the binding of the other partner and also ensure the stability of the linkage in the face of fluctuations in local concentration. A kinetic barrier to binding might also prevent inappropriate assembly if the ligand concentration is unstable.

The concept of vinculin as an integrator of spatial localization signals might apply broadly to the control of junctions and other subcellular structures. Understanding these assemblies will require measurement of the local concentrations of the various components, as well as the thermodynamic and kinetic parameters of their interactions. Only certain combinations of components present at particular times and concentrations might provide sufficient energy to allow assembly to proceed. In this way, diverse signals that influence spatial and temporal localization can be integrated to produce functional subcellular assemblies. ■

William I. Weis is in the Departments of Structural Biology and of Molecular and Cellular Physiology, Stanford University School of Medicine, 299 Campus Drive West, Stanford, California 94305-5126, USA.
e-mail: bill.weis@stanford.edu

1. Bakolitsa, C. *et al.* *Nature* **430**, 583–586 (2004).
2. Borgon, R. A., Vornrhein, C., Bricogne, G., Bois, P. R. J. & Izard, T. *Structure* **12**, 1189–1197 (2004).
3. Izard, T. *et al.* *Nature* **427**, 171–175 (2004).
4. Izard, T. & Vornrhein, C. *J. Biol. Chem.* **279**, 27667–27678 (2004).
5. Johnson, R. P. & Craig, S. W. *Nature* **373**, 261–264 (1995).
6. Johnson, R. P. & Craig, S. W. *J. Biol. Chem.* **269**, 12611–12619 (1994).
7. Kroemker, M., Rudiger, A. H., Jockusch, B. M. & Rudiger, M. *FEBS Lett.* **355**, 259–262 (1994).
8. Gilmore, A. P. & Burridge, K. *Nature* **381**, 531–535 (1996).
9. Weekes, J., Barry, S. T. & Critchley, D. R. *Biochem. J.* **314**, 827–832 (1996).
10. Steimle, P. A., Hoffert, J. D., Adey, N. B. & Craig, S. W. *J. Biol. Chem.* **274**, 18414–18420 (1999).
11. Bass, M. D. *et al.* *Biochem. J.* **362**, 761–768 (2002).
12. Bakolitsa, C., de Pereda, J. M., Bagshaw, C. R., Critchley, D. R. & Liddington, R. C. *Cell* **99**, 603–613 (1999).

Evolutionary biology

Oceans of bacteria

Stephen Giovannoni

Cloning microbial genes from natural environments has revealed a surprising amount of diversity. In understanding how microorganisms function in ecosystems, how much of this diversity really matters?

On page 551 of this issue, Acinas and co-workers¹ provide fresh fuel for the debate about what constitutes a bacterial species. They have analysed microbial diversity in nature using the polymerase chain reaction to clone and sequence a certain gene, the 16S ribosomal RNA gene, a widely used way of detecting organisms that defy cultivation in the laboratory^{2,3}.

Their subject, sea water, is perhaps the most intensively studied microbial habitat on the planet. What is new in this study is the use of techniques to reduce sequence artefacts, and the emphasis on the fine structure of evolution. When Acinas *et al.* plotted sequence similarity against the number of sequence comparisons, they observed a sharp discontinuity. In other words, when

they reduced the error in their sequences, the resulting evolutionary trees, like bonsai, showed distinct signs of pruning. They suggest that the bushy tips of their trees ('microdiverse clusters') may represent populations of cells that share similar ecological niches and adaptations, and therefore could be regarded as natural taxonomic units — species or 'ecotypes' of species.

At the root of this issue is the observation that copies of microbial genes, recovered by cloning from nature, are rarely identical, which can lead to high estimates of diversity. Earlier this year, Venter and colleagues⁴ published more than one billion base pairs of genomic sequence data from microorganisms inhabiting surface waters of the Sargasso Sea — equivalent to approximately

775 complete microbial genomes. The technique they used, shotgun sequencing, involves the assembly of sequence fragments into putative genomes. Frustrated by the difficulty of this procedure, Venter *et al.* argued that the high genomic DNA diversity they encountered was evidence that their samples were populated by at least 1,800 species. Their argument was based on extrapolation from a common rule-of-thumb, which classifies organisms that are more than 3% different in 16S rRNA sequence as different species. Sequences in the microdiverse clusters produced by Acinas *et al.* are more than 99% similar, thus falling within a species according to the 3% rule.

To take a particular example, Venter's team found that genes from one of the bacterial groups studied by Acinas *et al.*, SAR11 (*Pelagibacter*), accounted for 380 of the 1,412 16S rRNA genes they recovered. But the largest SAR11 fragment they could piece together was relatively small (about 21,000 base pairs), and even this sequence was not often repeated in the clone library, suggesting very high genomic DNA diversity within this group. In contrast, Acinas *et al.* found that most SAR11 rRNA genes from their samples could be placed in four or five microdiverse clusters, implying relatively limited diversity.

It is unclear how these data can be reconciled. Part of the problem may lie with the 3% rule and similar guidelines that do not take into account that most sequence change is caused by the clock-like accumulation of neutral sequence variation — that is, variation that has neither a positive nor a negative selective effect. Because new species arise randomly over time, high neutral sequence diversity within an ancient species could cause it to be construed as many species, at least in theory.

The bacterial-species issue may seem esoteric, but it is now assuming prominence as gene sequences from nature are applied to understand the global role of bacteria in biogeochemical processes. The question is, does each small branch of a gene tree represent an organism that plays a unique role in nature, or do the bushy tips of the trees represent sets of organisms (species or ecotypes of species) that essentially play the same role? The second explanation is obviously attractive because it reduces the complexity of modelling the various processes.

For most eukaryotic organisms, including plants, animals, fungi and protists, a species is defined as an interbreeding population. But bacteria don't have sex. Instead they use 'parasexual' processes, which, although much less efficient, accomplish the same thing, and are far more likely to result in the acquisition of DNA from unrelated species. In the 1980s, Sonea and Panisset⁵ argued that the discoveries of molecular biology render the bacterial-species concept

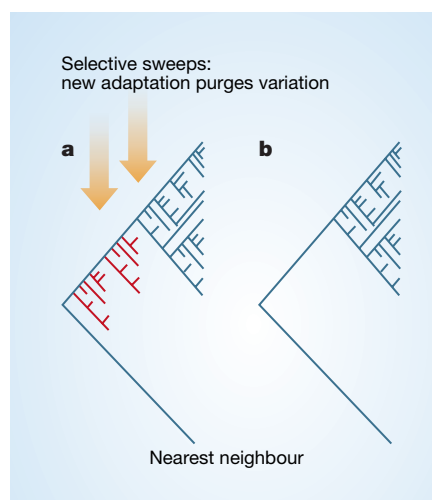


Figure 1 Selective sweeps and bonsai trees.

a, The data of Acinas *et al.*¹ support a model of microbial evolution in which selective sweeps — the spread of advantageous new genes — purge neutral genetic variation. b, The result is loss of diversity (red segments), and evolutionary trees that look like bonsai with bushy clusters of branches connected to their nearest neighbours by long internal segments. According to this view, the bushy tips of the trees represent the accumulation of neutral variation within microbial populations that occupy the same niche and have similar biogeochemical activity.

obsolete, and that bacteria are indeed a superorganism with a common gene pool. This idea received further support with the onset of the age of genomics, when it was revealed that most microorganisms are composites of genetic information from many sources⁶. So much DNA has been traded over the eons that almost every cell is a patchwork quilt.

But a big question remains: is the rate of this horizontal gene exchange in bacteria rapid enough to disrupt the emergence of distinct sets of physical characteristics (phenotypes) with superior competitive advantages^{7,8}? Acinas and colleagues' results are not what Sonea and Panisset's theory predicts: the observed pattern of rRNA gene evolution is not at all random, and strongly suggests that selection acts to create populations of cells that share the same niche, and to all intents and purposes might be regarded as species.

This will not be much of a surprise to microbial systematists, who, in the tradition of Linnaeus, have continued to add to and refine the list of named bacterial species. These microbiologists have not acted out of blind adherence to tradition, but out of respect for a common observation. Regardless of the evolutionary processes at work, microbiologists know that bacterial isolates in culture can be grouped into clusters of strains that are recognizable by their phenotypic traits.

The patterns of evolution observed by

Acinas *et al.* fit an evolutionary model advanced by Cohan⁹, which is based on simple evolutionary assumptions and population-genetic theory (Fig. 1). Cohan argued that evolutionary trees would reveal the history of microbial species by showing when valuable mutations arose that allowed descendants of that cell to dominate, essentially by growing faster, or by evading predation better, than competing cells. In this view, microbial cells in nature are highly adapted and constrained by selection, with any new and better cell rapidly proliferating and eliminating lesser competitors. It is this process, the emergence of new and better cells, that causes periodic selection, also known as selective sweeps, to prune the inner branches from trees.

It seems paradoxical that an apparently successful and competitive set of organisms such as the SAR11 group should also exhibit high diversity in their genomic DNA: one might expect that keen selection would truncate diversity. This is particularly the case for populations with large effective population sizes, which Kimura¹⁰ predicted would be less subject to the process known as genetic drift and therefore more able to 'fix' genetic variants that confer a small selective advantage. The solution to this paradox may be that very large populations that have not been through recent episodes of purifying selection can maintain large reservoirs of neutral genetic variation. If so, then at least within ecotypes of microbial plankton, one would expect to find a core set of genes conferring a relatively conserved phenotype.

Larger and more accurate 16S rRNA data sets than that used by Acinas *et al.* are already available from environmental sequencing studies⁴. Moreover, these data sets include flanking DNA sequences that may tell us a great deal about genome evolution. Such additional information provides extraordinarily detailed snapshots of evolution in action that are likely to reshape our understanding of the forces that control microbial diversity. The debate about the origin of microbial species shows no sign of ending, but it is certainly heating up.

Stephen Giovannoni is in the Molecular and Cellular Biology Program, Department of Microbiology, Oregon State University, Corvallis, Oregon 97331, USA.

e-mail: steve.giovannoni@orst.edu

1. Acinas, S. G. *et al.* *Nature* **430**, 551–554 (2004).
2. Rappe, M. S. & Giovannoni, S. J. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
3. Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. *Nature* **345**, 60–63 (1990).
4. Venter, J. C. *et al.* *Science* **304**, 66–74 (2004).
5. Sonea, S. & Panisset, M. *A New Bacteriology* (Jones & Bartlett, Boston, MA, 1983).
6. Nelson, K. E. *et al.* *Nature* **399**, 323–329 (1999).
7. Doolittle, W. F. *Trends Cell Biol.* **9**, M5–M8 (1999).
8. Daubin, V., Moran, N. A. & Ochman, H. *Science* **301**, 829–832 (2003).
9. Cohan, F. M. *Syst. Biol.* **50**, 513–524 (2001).
10. Kimura, M. *Nature* **217**, 624–626 (1968).

Nuclear physics

Not-so-magic numbers

David Warner

When a nucleus has a 'magic' number of neutrons or protons, it is particularly stable. But it seems that for exotic nuclei, with large numbers of neutrons relative to protons, these magic numbers can change.

Most people are familiar with the concept of an atom as a mini Solar System, with electrons orbiting a central nucleus. Not so many realize that the neutrons and protons within the nucleus behave in the same way. Despite the lack of a central nuclear 'Sun', the complex forces between the constituents of the nucleus produce an average potential that is central and that results in well-defined orbits for the neutrons and protons. Moreover, just as certain atoms are chemically more stable than others because they correspond to the filling of major electron shells (the noble gases, for instance), so nuclei with particular 'magic' numbers of neutrons and/or protons — 2, 8, 20, 28, 50, 82, and so on — have enhanced stability. These magic numbers have been known and understood^{1,2} for more than 50 years — until recently, when studies of extremely exotic nuclear species began to indicate that the magic numbers can change. In *Physical Review Letters*, Schiffer and colleagues³ report results that dramatically reveal the onset of this effect in isotopes of tin. They suggest that a decreasing nuclear 'spin-orbit' interaction is to blame.

The challenge to the basic assumption of the permanence of nuclear magic numbers has developed through studies of nuclei that, in terms of their numbers of protons and neutrons, are far from the region of stable, naturally occurring isotopes. It seems that the gaps in the sequence of nuclear energy levels that give rise to the extra stability of magic-number nuclei are actually quite fragile and shift with changing nucleon (proton or neutron) numbers. So far, the evidence has been indirect and limited to lighter elements with large neutron excesses. For such elements, the properties associated with enhanced stability (spherical shape, difficult to excite) that would be expected for nuclei with magic neutron numbers 20 or 28, for example, are not always observed; in some cases the shell effects have vanished, only to reappear at different neutron numbers.

One feature that produces very different magic numbers in stable nuclei and atoms is the strength of the spin-orbit interaction, a force that depends on the direction of the intrinsic spin of a particle (electron or nucleon) relative to that of its orbital motion. This interaction is much larger in the nucleus than it is for electrons, with the result that when the directions of spin and orbital angular momentum are aligned, the orbits are

pushed down in energy relative to orbits for which spin and orbital angular momentum are opposed. Theoretical studies^{4,5} attempting to solve the mean-field problem in nuclei (whereby the interactions between all the nucleons in the nucleus give rise to the average shell-model potential) have suggested that an increasing excess of neutrons may lead to a steady decrease in the strength of this spin-orbit interaction and that this may contribute to a radical change in the ordering of energy levels in the potential. The experiment carried out by Schiffer *et al.*³, at the tandem Van de Graaff accelerator at Yale University, has provided the first empirical evidence that this is indeed the case.

Schiffer and colleagues attacked the problem by studying a range of tin isotopes undergoing a nuclear reaction that transferred a single proton into specific vacant orbits of a tin nucleus. They directed a beam of 40-MeV α -particles (helium nuclei, ${}^4\text{He}$) onto a target of a single isotope of tin, and detected the tritons (tritium nuclei, ${}^3\text{H}$) that emerged whenever a proton had been transferred from an α -particle to a tin nucleus. The momentum of the tritons was measured using a magnetic spectrograph so that the populations of different energy levels in the final nucleus — corresponding to the proton entering different orbits of the shell model — could be distinguished. Particular care was taken to ensure that the absolute probability, or cross-section, for populating each state was accurately determined; these cross-sections were also measured as a function of angle relative to the direction of the beam.

The orbits of interest are those with the highest values of the orbital angular momentum, l . These orbits have the largest spin-orbit interactions and can contain the largest number of nucleons, so changes in their position have a dramatic effect on the magic numbers. The ideal experiment would have been to locate and track the relative positions of both members of a spin-orbit doublet — that is, two states with total angular momentum corresponding to the parallel and anti-parallel coupling of the spin to a specific l -value. Unfortunately, in nuclei this is not possible for higher values of l , because the splitting between the two members of the doublet becomes so large that both states cannot be observed in the same nucleus. The experimenters decided instead to focus on comparing the upper member of the highest- l doublet in the shell (with anti-

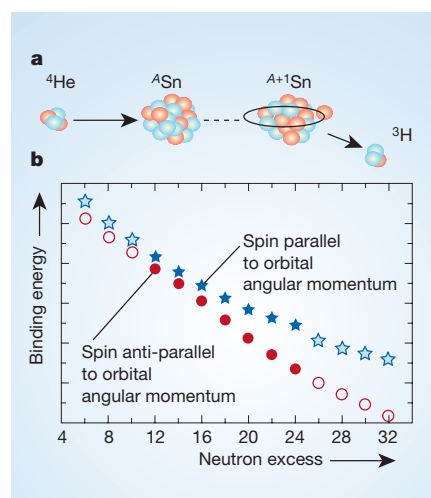


Figure 1 Neutron excess and the energy of states. **a**, Firing a beam of α -particles (${}^4\text{He}$) at a tin target causes the transfer of a proton to a tin nucleus of mass A (${}^A\text{Sn}$), and a triton (${}^3\text{H}$) emerges. The proton can enter one of two nuclear orbits with high angular momentum: one in which the proton's spin is parallel to its orbital angular momentum, the other in which the spin is opposed to its orbital motion. **b**, Schiffer *et al.*³ have measured the binding energies of these two orbits as a function of the neutron excess — that is, the number by which neutrons exceed the 50 protons in a tin nucleus. The separation of the states increases with increasing neutron excess, suggesting that the magnitude of the spin-orbit interaction is changing. (Filled symbols are Schiffer and colleagues' data, open symbols represent data not yet confirmed by transfer reactions; **b** derived from ref. 3.)

parallel spin) with the lower, spin-aligned member of the highest- l doublet from the shell above (which is pushed down by the spin-orbit interaction).

The results are shown in Figure 1, and the steadily increasing separation of the two states is obvious. The accurate experimental determination of the cross-sections means that these states can be associated unambiguously with particular proton orbits, and hence this is the first direct evidence in heavy nuclei of the changing relative energy of these high- l proton orbits with increasing neutron excess. Moreover, this association can now be extended with confidence to analogous states in the most exotic tin nuclei. The authors present a similar analysis of nuclei with one neutron outside the 82-neutron closed shell — albeit with less quantitative backing — that suggests a similar trend exists for two higher-energy spin-opposed and spin-parallel neutron states, again consistent with a decreasing spin-orbit interaction with increasing neutron excess.

Understanding the origin of the shifting nuclear shell structure is not just a challenge for nuclear physicists. The nuclei that make up our world have been manufactured in

nuclear interactions in the stars. Many of these processes of nucleosynthesis take place far from conditions of stability, on timescales short enough to prevent an unstable nucleus decaying before it is involved in another reaction. Thus, the properties of highly exotic nuclear species are central to the problem, and changes in the nuclear magic numbers will have a profound impact on the reaction rates that determine the formation of elements in the stars. ■

David Warner is in the Surface and Nuclear Division, Daresbury Laboratory, Daresbury, Warrington, Cheshire WA4 4AD, UK.
e-mail: d.warner@dl.ac.uk

1. Mayer, M. G. *Phys. Rev.* **75**, 1969–1970 (1949).
2. Haxel, O., Jensen, J. H. D. & Suess, H. E. *Phys. Rev.* **75**, 1766 (1949).
3. Schiffer, J. P. *et al.* *Phys. Rev. Lett.* **92**, 162501 (2004).
4. Dobaczewski, J., Hamamoto, I., Nazarewicz, W. & Sheikh, J. A. *Phys. Rev. Lett.* **72**, 981–984 (1994).
5. Lalazissis, G. A., Vretenar, D., Pöschl, W. & Ring, P. *Phys. Lett. B* **418**, 7–12 (1998).

Self-assembly

Towards precision micelles

Dennis E. Discher and Randall D. Kamien

Detailed imaging reveals the structure of a spherical ‘micelle’, self-assembled from cone-shaped molecules, and marks progress towards mimicking the natural assembly skills of biological systems.

Cells synthesize and assemble molecules, even complex machines, of exacting but non-symmetric structure. The precision of the building-blocks and the checks and rechecks of assembly ultimately establish the variety and complexity of cellular function, replication and division. Could man-made molecules mimic those in the cell and also form precision materials? Apparently so. In *Angewandte Chemie International Edition*, Kellermann *et al.*¹ provide a convincing demonstration in their work on amphiphilic dendro-calixarenes.

Calixarenes are cone-shaped molecules; ‘amphiphilic’ describes their tendency to self-assemble into aggregates. The dendro-calixarenes in Kellermann and colleagues’ study have T-shaped dendrimer, or branching, heads, and these coordinate the self-assembly of exactly seven molecules into a micelle with a non-polar core and a hydrophilic exterior (Fig. 1). This is the first structural characterization of a micelle on the molecular scale, made possible by high-resolution imaging techniques that include reconstructions from cryogenic transmission electron microscopy (or ‘cryo-TEM’). Such techniques will clearly soon be playing a major role in the characterization and synthesis of ordered soft materials.

The symmetry of these calixarene micelles is remarkable: two pairs of T-shapes bind through specific lock-and-key interactions, and another three molecules complete the ensemble (Fig. 1c). What emerges is a near-spherical micelle with C_2 symmetry — that is, the micelle is symmetric under rotation through an angle of π radians. The micelles are also surprisingly stable and persist even after removal of the solvent; such behaviour is consistent with the existence of strong interactions that dictate order beyond assembly. Coordination between the heads of the

calixarenes seemingly opposes disassembly and reorganization.

Although this appears to be the first and most highly ordered example of a persistent spherical micelle, a large family of structures with extended and highly stable cylindrical morphologies has been generated using amphiphilic peptides, or short stretches of amino acids; the peptides contribute to β -sheet structures along the length of a cylinder². These nanofibres are showing promise

as self-assembled scaffolds for cells in various contexts. Another example³ is based on amphiphilic block copolymers (composed of covalently linked polymer units). These, through metal-coordinating monomers, seem to crystallize within their cylindrical core⁴. Self-assembled nanowires for molecular electronics are a possibility with such linear architectures⁵.

However, to obtain spheres as Kellermann *et al.*¹ have done, the coordinating interactions must not be linear. In the T-calixarene micelles, the unique arrangement of seven amphiphiles not only breaks the symmetry of the micelle but also suggests that the entire structure should be polar. This in turn should make it possible to align the micelles in fields (magnetic or electric) for more accurate characterization and possibly for use as macromolecular ferrofluids or electrorheological fluids. It might also allow higher-order assembly, although probably under different solution conditions, to give chains that are akin to biofilaments such as actin, which is integral to the structure of the cell⁶.

What sort of interactions could create the highly asymmetric structure within the T-calixarene micelle? When seven spheres are brought together by a surrounding interface, they tend to form an arrangement that has a different symmetry⁷ from the C_2 symmetry of these micelles. This suggests that simple packing of the T-calixarenes is

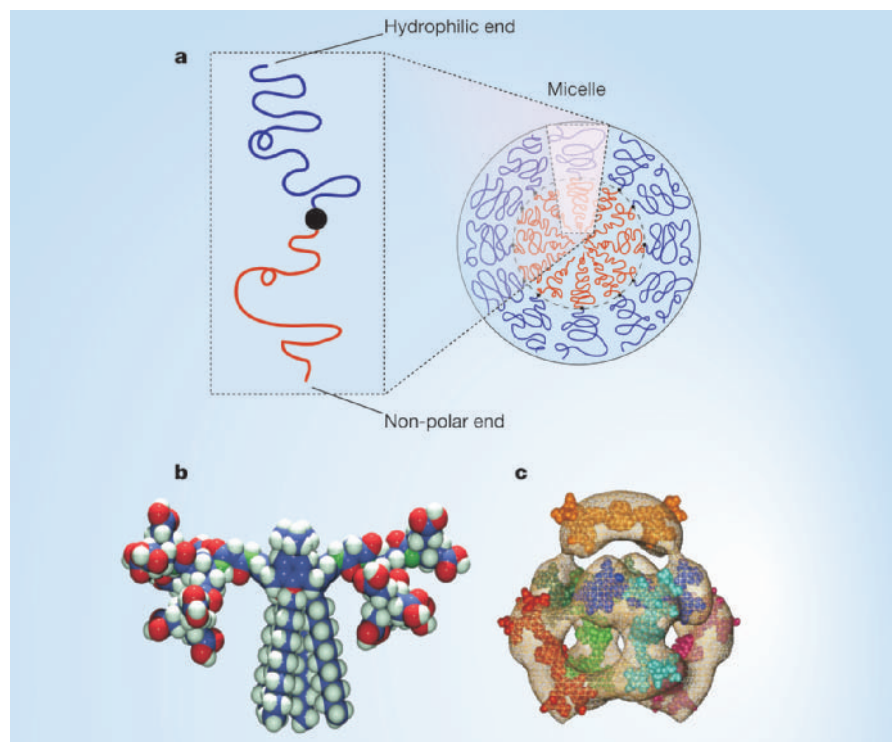


Figure 1 The seven-molecule spherical micelle. **a**, Amphiphilic molecules, which have a hydrophilic end and a non-polar end, can arrange themselves into a micelle, with all the hydrophilic ends at the outermost edge. **b**, Kellermann *et al.*¹ have created micelles from T-shaped molecules composed of cone-like calixarene molecules with dendritic heads. The low-energy conformer is shown here. **c**, High-resolution imaging reveals a curious structure of two pairs of intercalated T-calixarenes, with three additional molecules ‘clipped’ around them. (**b** and **c** reproduced, with permission, from ref. 1.)



100 YEARS AGO

The best angle of traction on a rough or irregular surface is at an upward inclination to its general slope. This upward slanting pull can be applied to a two-wheeled vehicle, and to the fore-wheels of any vehicle, but not to the hind wheels — especially if they are far away. Consider, further, the summit of a hill, and let a waggon be so elongated that its hind wheels are still ascending while the horse is descending: his pull is exerted at a very bad angle on this part of the load... I should like to take this opportunity of saying that whether the traditional heavy draught of a long-bodied carriage is well founded or not, I am convinced that the ordinary hansom cab is badly balanced, and that a horse would be better with some load on his back, except when descending a hill... Nothing can be worse than a constant upward pressure on the chest of a horse: a pressure which at present automatically increases on an up grade, thus tending to deprive the animal of part of his own weight, on the existence of which the efficacy of every locomotive depends.

Oliver Lodge

From *Nature* 28 July 1904.

50 YEARS AGO

The Papers of Wilbur and Orville Wright. These two volumes give for the first time a full account by the Wright brothers of the history of the solution of the problem of human flight. In the introduction it is explained that they fully intended to write up the results of their work for publication, but that they became so preoccupied with their experiments and later with the marketing of their machines that they never did so. After Wilbur's death in 1912, Orville had the same good intention, but never carried it out. After his death in 1948, the voluminous correspondence of the brothers and their note-books eventually came into the possession of the United States Library of Congress, and it is from this material that the present volumes have been compiled... The second volume contains later experimental work; but is largely taken up with the efforts made by the brothers, sometimes separately and sometimes together, to obtain some reasonable financial recompense for all their work by selling their invention to governments or private companies, a matter in which they met with no small difficulty.

From *Nature* 31 July 1954.

not responsible for their arrangement. Note that if the crossed pairs of interacting T-calixarenes are viewed as single entities, the arrangement still lacks inversion symmetry. It seems instead that the key to the micelle assembly is a non-planar linking of the T-shaped pairs, apparently mediated through coordinated hydrogen bonding. Once the pairs interlock, they become a curved building-block that imposes spherical geometry. The net effect is to exclude the formation of linear micelles or even membranes.

By adding moieties that prefer the local chemistry of the interlocking T-calixarenes, it should be possible to exploit the micelle's net polarity and make asymmetrically functionalized micelles with geometric accuracy. For instance, the four proteins that assemble into the histone octamer are subsequently involved in forming the complex three-dimensional structure that is the

chromosome — an example that might prove fruitful for better understanding the physical principles in the packing of DNA. Other applications might include hierarchies of aggregates: precision micelles inside larger, perhaps less-ordered aggregates could deliver a targeted regimen of several therapeutic drugs.

Dennis E. Discher and Randall D. Kamien are in the Laboratory for Research on the Structure of Matter, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

e-mails: discher@seas.upenn.edu

kamien@physics.upenn.edu

1. Kellermann, M. *et al.* *Angew. Chem. Int. Edn* **43**, 2959–2962 (2004).
2. Hartgerink, J. D., Beniash, S. & Stupp, E. I. *Science* **294**, 1684–1688 (2001).
3. Discher, D. E. & Eisenberg, A. *Science* **297**, 967–973 (2002).
4. Resendes, R. *et al.* *Chem. Eur. J.* **7**, 2414–2424 (2001).
5. Percec, V. *et al.* *Nature* **419**, 384–387 (2002).
6. Pollard, T. D. *Nature* **422**, 741–745 (2003).
7. Manoharan, V. N., Elsesser, M. T. & Pine, D. J. *Science* **301**, 483–487 (2003).

Molecular biology

Cohesins slip sliding away

Karen E. Ross and Orna Cohen-Fix

Cohesin complexes have a central role in cell division, mediating the association between sister chromosomes. It now seems that cohesin binding is dynamic, adapting to changes in gene transcription.

Dividing cells face the monumental task of providing each progeny cell with a complete and accurate copy of the genome. During mitosis, the basic cell-division process, cells need to ensure that they send only one copy of every chromosome to each of the daughter cells. Cohesion between the two sister chromatids produced by chromosome duplication is essential to this process: following duplication, the sister chromatids are physically linked by an evolutionarily conserved, ring-shaped protein complex known as cohesin¹.

Cohesin binding serves three functions: it allows cells to keep track of the two sister chromatids, thereby ensuring their proper segregation in mitosis; it facilitates attachments between chromosomes and the spindle, the apparatus that segregates chromosomes into the two daughter cells; and it facilitates DNA repair by recombination. Cohesin-mediated cohesion between sister chromatids also has a vital role during meiosis, the specialized cell division that produces the reproductive cells.

It is not known how cohesins link sister chromatids together; cohesin might bind to chromosomes directly, bridging the two sister chromatids, or, given its ring-like structure, it might form a ring that encircles one or both sisters². Regardless of the mechanism, the presence of cohesin is likely to cause a logistical problem: when behemoth enzyme complexes such as RNA polymerase, which

transcribes genes into RNA, slide along cohesin-bound chromosomes, they have to contend with these complexes in their path. Lengronne *et al.* (page 573 of this issue³) and Glynn *et al.* (writing in *PLoS Biology*⁴) have created high-resolution maps of cohesin-binding sites throughout the genome of budding yeast, and have thereby gained tantalizing insight into how the transcription machinery and cohesins coexist.

The importance of cohesin has generated great interest in identifying cohesin-binding sites on chromosomes. DNA-binding proteins often show a preference for particular DNA sequences, but, curiously, no such sequences had been identified for cohesin. Previous work revealed that cohesins are present at centromeres, the sites on chromosomes that mediate attachments to the spindle, and along chromosome arms, where they are bound with an average spacing of 10 kilobases^{5–7}. Chromosome-arm binding sites tend to be in regions between genes — intergenic regions — that are rich in adenine (A) and thymine (T) nucleotides. However, these previous studies did not examine large portions of the genome at high resolution. Therefore, Lengronne *et al.*³ and Glynn *et al.*⁴ set out to create a more complete picture of cohesin-binding sites, hoping to uncover the underlying principles that govern cohesin localization.

Both groups generated their cohesin maps using a technique called ChIP-chip,

for chromatin immunoprecipitation followed by analysis on a microarray, or 'gene chip' (Fig. 1). Yeast cells were first treated with chemical crosslinkers to attach cohesins covalently to their chromosomal binding sites. The DNA was then sheared, and antibodies against one of the cohesin subunits were used to precipitate cohesin complexes and their associated DNA. This DNA was then amplified, labelled and incubated with a microarray consisting of DNA fragments of known sequence that represented much of the yeast genome. The labelled DNA bound to matching sequences on the array, thereby identifying the sequences corresponding to cohesin-binding sites in living cells.

The new maps for cohesin-binding sites revealed that, along chromosome arms, cohesins did not bind randomly to intergenic regions. Rather, cohesin-site preference was influenced by the orientation of the neighbouring genes. Genes have a defined polarity: they are transcribed from one end

(head) to the other (tail). Neighbouring genes can be found in various arrangements relative to each other: head to head, head to tail, or tail to tail. Of more than 1,000 cohesin-association sites that map to chromosome arms, the vast majority occur between genes that are arranged tail to tail, where transcription of the two genes converges (Fig. 2a). It might be that a specific feature of the nucleotide sequence or a particular chromatin modification attracts cohesins to these zones of convergent transcription. However, an intriguing possibility raised in both reports is that RNA polymerase, as it moves along the chromosome, acts like a snowplough and shoves cohesins along the DNA, causing them to pile up where RNA polymerases travelling in opposite directions meet.

Consistent with this model, Lengronne *et al.*³ and Glynn *et al.*⁴ show that when transcriptional activity changes, the pattern of cohesin-binding sites also changes. Cohesins appear on genes whose transcription is turned off and disappear from genes whose transcription is turned on (Fig. 2b). The observed relocation of cohesin was rapid, in one case occurring within 15 minutes of the start of transcription³. In this experiment, an increase in transcription was associated not only with the loss of cohesin from the gene-coding region, but also with accumulation of cohesin near the gene's tail (Fig. 2b). This rearrangement could come about in one of two ways: cohesin could dissociate and rebind, or it could slide along the DNA, piling up at the end of the gene as a result of the movement of the RNA polymerase. The second possibility is attractive given cohesin's ring structure, but it would mean that this rearrangement is independent of the factors necessary for loading cohesin onto DNA⁸ — which has yet to be proven.

Not all cohesin was found in regions of convergent transcription. Glynn *et al.*⁴ examined several cohesin-binding sites within genes and found them to be AT-rich, suggesting that along chromosome arms AT-richness and convergent transcription may be two independent determinants of cohesin binding. Consistent with previous reports, the region surrounding the centromere was also thickly coated with cohesin. Binding of cohesin to this region depends on chromosome-associated proteins that mediate the attachment of the chromosome to the spindle⁹. In animal cells, cohesin is also found near centromeres, and in this case it is recruited by factors in heterochromatin, a specialized, highly compact chromatin structure¹⁰.

Cohesins are also associated with zones of convergent transcription in fission yeast³, but it remains to be seen whether the same is true in multicellular organisms. Unlike yeast genes, which are generally small, the genes of multicellular organisms can be huge, and zones of convergent transcription may be

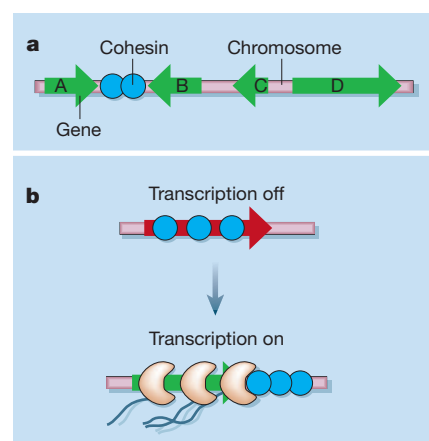


Figure 2 Cohesins associate with zones of convergent transcription. **a**, Genes occur in several orientations along the chromosome; green arrows indicate the direction of transcription. Genes A and B are arranged tail to tail; genes B and C are head to tail; and genes C and D are head to head. Cohesins are generally found in the spaces between genes that are arranged tail to tail. **b**, Transcription leads to cohesin rearrangement. When transcription of a gene switches from off (red) to on (green), cohesins are lost from the coding region and instead associate near the end of the gene. RNA polymerase is shown in beige and the newly synthesized RNAs are shown as thin black lines.

too widely spaced to provide sister chromatids with sufficient cohesion. Interestingly, vertebrates have two types of cohesin complex with different subunit compositions¹, raising the possibility that they differ in their binding-site preference. Thus, the rules that govern cohesin localization in other organisms await further investigation.

The creation of a genome-wide map of cohesin-binding sites in budding yeast suggests that sister chromosome cohesion is much more dynamic than previously imagined, adapting to the immediate transcriptional needs of the cell. Cohesin rearrangement could occur by a surprisingly simple mechanism — in the face of RNA polymerase, cohesins may go slip sliding away.

Karen E. Ross and Orna Cohen-Fix are at the Laboratory of Molecular and Cellular Biology, National Institute of Diabetes & Digestive & Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA.

e-mails: kross@helix.nih.gov

ornacf@helix.nih.gov

1. Haering, C. H. & Nasmyth, K. *BioEssays* **25**, 1178–1191 (2003).
2. Campbell, J. L. & Cohen-Fix, O. *27*, 492–495 (2002).
3. Lengronne, A. *et al.* *Nature* **430**, 573–578 (2004).
4. Glynn, E. F. *et al.* *PLoS Biol.* **2**, e259 (2004).
5. Megee, P. C., Mistrot, C., Guacci, V. & Koshland, D. *Mol. Cell* **4**, 445–450 (1999).
6. Blat, Y. & Kleckner, N. *Cell* **98**, 249–259 (1999).
7. Tanaka, T., Cosma, M. P., Wirth, K. & Nasmyth, K. *Cell* **98**, 847–858 (1999).
8. Ciosk, R. *et al.* *Mol. Cell* **5**, 243–254 (2000).
9. Weber, S. A. *et al.* *PLoS Biol.* **2**, e260 (2004).
10. Bernard, P. & Allshire, R. *Trends Cell Biol.* **12**, 419–424 (2002).

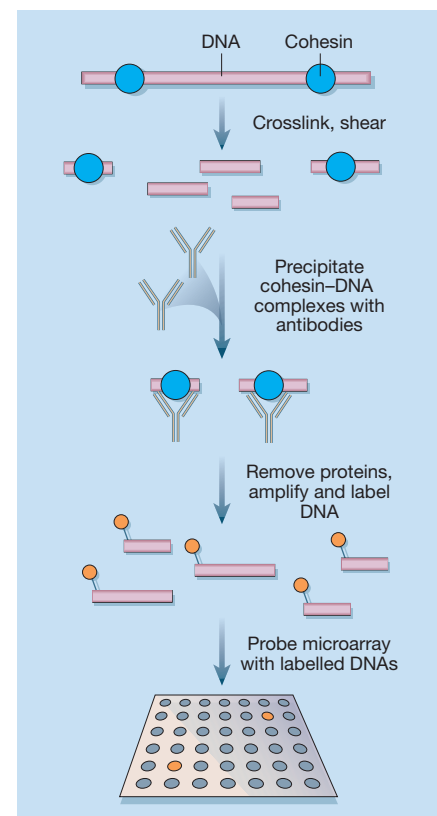


Figure 1 The ChIP-chip technique used by Lengronne *et al.*³ and Glynn *et al.*⁴. Chemical crosslinkers link cohesins (blue) to their chromosomal binding sites. The DNA (pink) is sheared into fragments of 0.2–1 kilobases, and antibodies against one of the cohesin subunits (Y-shaped structures) precipitate the cohesin complexes and their associated DNA. This DNA is amplified, labelled and incubated with a microarray containing known yeast DNA sequences at designated spots, represented by circles. The orange circles indicate places on the microarray where the precipitated DNA bound to a matching sequence.

Astronomy

Light-bending reveals stellar mass

Astrophys. J. (in the press); preprint at <http://arXiv.org/abs/astro-ph/0405124> (2004)

It's not easy to deduce a star's mass. At present, it is usually done indirectly, by using spectroscopic measurements to determine the star's radius and surface gravity. Direct mass measurements are possible for binary stars via their orbital parameters, but the only single star for which a truly direct mass measurement has been made is the Sun. That was done by looking at how the Sun bends light coming from a distant source: essentially the same measurement that Eddington made in 1919 during a solar eclipse, confirming Einstein's general theory of relativity.

Andrew Gould *et al.* have now used this same principle to measure the mass of a star 2,000 light years away in our Galaxy. This star acted as the lens for a microlensing event in 1993, denoted MACHO-LMC-5, when it crossed the path of a more distant star and focused that star's light. This observation, part of a project to look for dark-matter candidate objects called MACHOs by their gravitational-lensing effect, has only now been analysed in detail. The MACHO-LMC-5 lens is a dim red star with an apparent mass about one-tenth that of the Sun ($\pm 17\%$).

Philip Ball

Evolution

Dark prospects

Biol. J. Linn. Soc. **82**, 359–366 (2004)

With the realization that the sooty consequences of Europe's industrial revolution favoured the dark form of the peppered moth (*Biston betularia*) over its light-coloured counterpart (pictured side by side), the insect entered the textbooks as a stalwart example of natural selection. But it now seems that the dark form, called *carbonaria*, is nearing the end of the line.

L. M. Cook and colleagues analysed surveys of the frequencies of the pigmented and original typical forms, and of an

intermediate form called *insularia*. The surveys were carried out in Britain during the later years of the twentieth century, when numbers of *carbonaria* were declining in response to falling pollution levels. Cook *et al.* brought mathematics to bear on the surveys, to estimate the changing prospects of the different forms in various times and places.

The *carbonaria* form is in terminal decline, they conclude. But the same cannot be said for *insularia*, which should continue to be viable even in the post-industrial environment. What's more, although it has been suggested that heavy pigmentation can be triggered by metal salts in the larval diet, the authors believe that the results indicate that the origin of the different pigment patterns is entirely genetic.

Michael Hopkin

Analytical chemistry

How low can you go?

Anal. Chem. doi:10.1021/ac049657j (2004)

A sensitivity record for detecting trace amounts of molecules in extremely dilute solution has been shattered. Sunia A. Trauger *et al.* report that their mass-spectrometry approach can detect 800 yoctomoles (800×10^{-24} moles) of des-Arg⁹-bradykinin, a peptide commonly used as a sensitivity standard. That's equivalent to only 480 molecules, and represents a 50-fold greater sensitivity than previously obtained by mass spectrometry.

Using the technique known as desorption/ionization on silicon, or DIOS, the molecules are first adsorbed onto a porous silicon wafer, then desorbed and ionized by a laser pulse. This detaches the molecules from the silicon surface without breaking the complex structure apart. The charged peptide then flies through a conventional mass spectrometer, which calculates the mass of the molecule based on its 'time of flight'.

The authors' key breakthrough is the addition of fluorinated 'chemical tethers' to the silicon surface which help grab even the most elusive molecules from solution. The authors claim that this system could detect tiny amounts of compounds from

a complex mixture by tailoring the tether to bind selectively to the desired target molecule.

Mark Peplow

Genetics

First aid for flies

PLoS Biol. **2**, e239 (2004)

Wound healing is essential for an organism's survival, but the genetic signals that control the process are unclear. Michael J. Galko and Mark A. Krasnow have developed a fly model of larval wound healing and used it to elucidate key regulatory genes.

The duo stabbed fruitfly larvae with needles to create non-fatal puncture wounds. Normally, a scab forms and neighbouring epidermal cells spread across the wound to regenerate the skin. But when the JNK (for c-Jun N-terminal kinase) signalling pathway is inactivated, although scabs form, epithelial migration is blocked. Moreover, larvae with defects in a gene needed for the generation of crystal cells — a type of blood cell — could not form scabs properly.

The study hints that cellular aspects of wound healing are coordinated by multiple, separate genetic signals. Cellular and genetic parallels exist between wound healing in flies and in mammals, so the researchers speculate that the process might be an ancient response that diversified during evolution.

Helen Pilcher

Biological chemistry

Kinase connections

J. Am. Chem. Soc. doi:10.1021/ja048659i (2004)

Protein phosphorylation is a major mechanism for signalling within a cell. But it's no easy matter to trace the biochemical pathway back from a known phosphorylation site to the kinase enzyme that catalyses the phosphate transfer. Dustin J. Maly and colleagues have tackled this problem, and come up with a new method for identifying upstream kinases.

The authors set out to develop a way of linking biologically relevant substrate-kinase pairs. To ensure specificity for the desired phosphorylation site, the 'phosphoacceptor' amino acid (serine or threonine) is replaced with cysteine. The crosslinking reagent is an ATP analogue, which is designed to form a link in a kinase-catalysed reaction with both a lysine amino acid at the enzyme's active site and the cysteine of the modified substrate. As a proof of principle, Maly *et al.* demonstrate specific crosslinking of kinase and peptide partners, and reactivity of the crosslinker with a range of serine/threonine kinases.

The next steps will be to identify previously unknown kinase-substrate pairs, and to extend the method for use *in vivo*.

Joanne Kotz



J. MASON/ARDEA

Ground squirrel uses ultrasonic alarms

This rodent emits a high-frequency shriek as a warning that is inaudible to predators.

Part from echolocation and the pursuit of prey by bats¹, the function of ultrasound in animal communication is poorly understood². This is mainly because of the broad range of responses that it can evoke and the widely varied contexts in which it is used (for example, in rodents of the Muridae family it may indicate distress in infants or a sexual or predatory encounter in adults)³. Here we find that a purely ultrasonic signal is produced in the wild by a rodent of the Sciuridae family, Richardson's ground squirrel, and show that its function is to warn conspecifics of impending danger. To our knowledge, ultrasonic alarm calls have not previously been detected in any animal group, despite their twin advantages of being highly directional and inaudible to key predators.

Ground-dwelling squirrels produce audible (8 kHz) alarm vocalizations to warn others of danger. The call recipients benefit from improved detection of predators, and callers benefit through kin selection⁴. We studied alarm communication among Richardson's ground squirrels



Figure 1 Careful whisper: a ground squirrel making an ultrasound call as a covert signal of danger.

(*Spermophilus richardsonii*; Fig. 1) and noticed that whereas the motor behaviour of some of these animals (10 of 181 individuals exposed to a model predator⁵) was consistent with alarm calling, they produced only faint sounds of rushing air. These 'whisper' calls, which we observed in all our study populations, contain pure ultrasonic frequencies of around 50 kHz (Fig. 2a) and so constitute a previously undescribed vocalization by Richardson's ground squirrels⁶.

We recorded whisper calls from 15 free-living squirrels (for methods, see supplementary information). The mean ($\pm$ s.e.m.) sound-pressure level of calls was 66.8 ± 2.1 decibels at a mean ($\pm$ s.e.m.) distance from the squirrel of 0.49 ± 0.02 m. The mean ($\pm$ s.e.m.) duration and dominant frequency of the primary syllable were 225 ± 8 ms and 48.0 ± 2.3 kHz, respectively (for details, see supplementary information).

We investigated call function by broadcasting whisper calls and three control calls (these were background noise, a pure tone that matched the whisper call's dominant frequency and an audible call) to recipient free-living squirrels at a site that was 60 km from the recording site. Receiver vigilance was scored (for methods, see supplementary information) and compared among treatments. It was found that the animals spent significantly more of their time on vigilant behaviour in response to the whisper calls and audible control than in response to background noise (Fig. 2b); however, responses to whisper calls were qualitatively different from responses to audible signals. The increased vigilance recorded in response to the pure-tone control (Fig. 2b) was not significantly different from that produced in response to whisper calls.

These results indicate that the function of whisper calls is to warn nearby conspecifics

of potential danger and that the dominant ultrasonic frequency is important. Audible calls evoked a more pronounced response than whisper calls, suggesting that whisper calls either convey less urgency than audible calls or that respondents react less conspicuously.

In addition to being inaudible to many rodent predators³, ultrasound (frequency above 15 kHz) differs from audible sound in that it attenuates rapidly and is highly directional⁷. These

apparent limitations as a warning signal may allow callers selectively to warn⁸ philopatric kin⁴ while remaining undetected by predators outside the signal's active space.

The attenuation and directional propagation of whisper calls need to be tested to determine whether they enable callers to remain cryptic and whether squirrels selectively beam calls to specific receivers. But selection is likely to favour the deployment of whisper calls under particular circumstances, such as those described here. These vocalizations function as a warning of approaching predators and, given their spectral characteristics, are likely to limit the audience and reduce the probability of detection by the predator.

David R. Wilson, James F. Hare

Department of Zoology, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada

e-mail: harejf@cc.umanitoba.ca

1. Simmons, J. A., Fenton, M. B. & O'Farrell, M. J. *Science* **203**, 16–21 (1979).
 2. Smith, W. J. *Am. Zool.* **19**, 531–538 (1979).
 3. Sales, G. & Pye, D. *Ultrasonic Communication by Animals* (Chapman & Hall, London, 1974).
 4. Sherman, P. W. *Science* **197**, 1246–1253 (1977).
 5. Hare, J. F. *Anim. Behav.* **55**, 451–460 (1998).
 6. Koepl, J. W., Hoffman, R. S. & Nadler, C. F. *J. Mamm.* **59**, 677–696 (1978).
 7. Pye, J. D. & Langbauer, Jr, W. R. in *Animal Acoustic Communication* (eds Hopp, S. L., Owren, M. J. & Evans, C. S.) 221–250 (Springer, Berlin, 1998).
 8. Witkin, S. R. *Condor* **79**, 490–493 (1977).
- Supplementary information accompanies this communication on Nature's website.
- Competing financial interests: declared none.

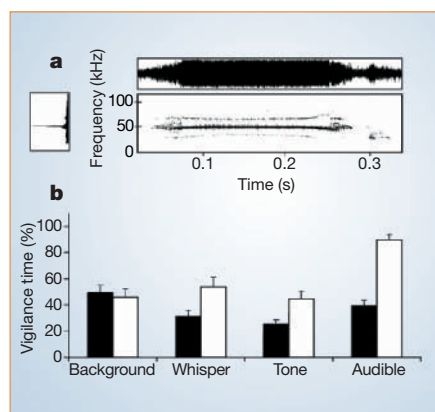


Figure 2 Ultrasonic 'whisper' calls by Richardson's ground squirrels and the response to them. **a**, Spectrogram (bottom) shows the call duration (238 ms) and dominant frequency (51.6 kHz) of the primary syllable; the signal intensity is represented along the time axis by the density of the grey scale. Inset left, power spectrum showing the intensity of individual frequencies (averaged across the signal). The time–amplitude window (top) shows the overall signal intensity relative to background noise. **b**, Proportion of time ($\pm$ s.e.m.) that squirrels ($n=19$) devoted to vigilant behaviour before (black bars) and during (white bars) the playback of whisper calls and of the three control calls. All experiments complied with the guidelines of the Canadian Council on Animal Care.

brief communications arising online

► www.nature.com/bca

Earth science: Role of fO_2 on fluid saturation in oceanic basalt

B. Scaillet & M. Pichavant (doi:10.1038/nature02814)

Reply: A. E. Saal, E. H. Hauri, C. H. Langmuir & M. R.

Perfit (doi:10.1038/nature02815)

Earth science

Role of f_{O_2} on fluid saturation in oceanic basalt

Arising from: A. E. Saal, E. H. Hauri, C. H. Langmuir & M. R. Perfit *Nature* **419**, 451–455 (2002)

Assessing the conditions under which magmas become fluid-saturated has important bearings on the geochemical modelling of magmas because volatile exsolution may profoundly alter the behaviour of certain trace elements that are strongly partitioned in the coexisting fluid¹. Saal *et al.*² report primitive melt inclusions from dredged oceanic basalts of the Siqueiros transform fault, from which they derive volatile abundances of the depleted mantle, based on the demonstration that magmas are not fluid-saturated at their eruption depth and so preserve the mantle signature in terms of their volatile contents. However, in their analysis, Saal *et al.*² consider only fluid–melt equilibria, and do not take into account the homogeneous equilibria between fluid species, which, as we show here, may lead to a significant underestimation of the pressure depth of fluid saturation.

For any basalt melt that is at fixed temperature and pressure in fluid-saturated conditions with known H_2O and CO_2 concentrations, the corresponding volatile fugacities, f_{H_2O} and f_{CO_2} , can be calculated³. The phase rule states that this in turn fixes the fugacities of all other C–O–H fluid species, including f_{O_2} (ref. 4). Figure 1a shows the covariation of the mole fraction of H_2O and CO_2 (X_{H_2O} and X_{CO_2}) in a C–O–H fluid calculated for various f_{O_2} at 1,200 °C and 400 bar (f_{O_2} expressed in log units relative to the solid buffer Ni–NiO, referred to here as NNO). It can be seen that at a very low mole fraction of H_2O ($X_{H_2O} < 0.05$), reduced fluids are poorer in CO_2 than oxidized ones: for instance, at $\Delta NNO = -2$ the mole fraction of CO_2 is 0.8, whereas at $\Delta NNO = -0.8$, it is 0.95. This is due to the progressive reduction of CO_2 into CO, which becomes significant at $f_{O_2} < \Delta NNO = -1$ (ref. 4).

Figure 1b shows the H_2O and CO_2 concentrations of basalt melts that coexist with fluids shown in Fig. 1a. Under oxidizing conditions ($f_{O_2} > \Delta NNO = -1$), the overall shape of the curve resembles the pattern of the curve when it is calculated by considering only fluid–melt equilibria². By contrast, for $f_{O_2} < \Delta NNO = -1$, the isobaric curve displays an asymmetric bell-shaped pattern characterized by a strong lowering of the melt CO_2 content at low H_2O . As already stated, this is the result of the reduction of CO_2 to CO at low f_{O_2} , CO being an insoluble species in silicate melts at low pressures⁵. The two curves merge at melt H_2O contents higher than 1 wt%, which shows that, for basalt melts with a higher meltwater content, the calculation of pressure for fluid saturation in the C–O–H system does not require an accu-

rate knowledge of their redox state — unlike H_2O -poor basalts, such as oceanic basalts².

The f_{O_2} of primitive melt inclusions at Siqueiros is at present not well constrained but is estimated to be around $\Delta NNO = -2$ (ref. 2), which would fall at the upper end of the range of f_{O_2} estimated for mid-ocean-ridge basalt⁶ (MORB). However, given the general inverse correlation between f_{O_2} and

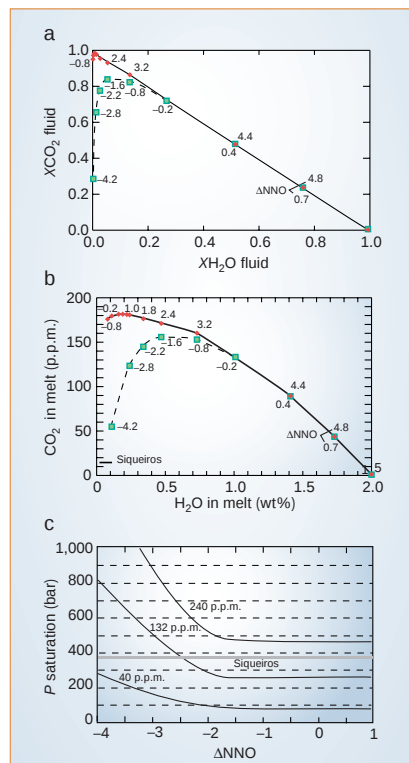


Figure 1 Effect of f_{O_2} on fluid speciation and fluid saturation in basalts. **a**, Covariation of X_{H_2O} and X_{CO_2} (where X_i is the mole fraction of species i) in a C–O–H fluid calculated for various values of f_{O_2} (numbers along each curve). The calculations were done by fixing f_{H_2} (either 0.01 bar or 1 bar, corresponding to red and green symbols, respectively) and f_{H_2O} , which allows us to calculate f_{CO_2} in the C–O–H system⁴. Once f_{H_2} and f_{H_2O} are fixed, f_{O_2} can be calculated through the equilibrium $H_2 + 0.5 O_2 = H_2O$. T , 1,200 °C; P , 400 bar. **b**, H_2O – CO_2 solubility diagram for a basalt at 1,200 °C and 400 bar and equilibrated with fluid compositions shown in **a**. For any given f_{H_2O} and f_{CO_2} set of values, the corresponding H_2O and CO_2 contents of the melt are computed according to ref. 3. The f_{O_2} is shown along each line in log units calculated relative to the solid buffer Ni–NiO. The Siqueiros bar shows the range of H_2O content of Siqueiros melt determined by Saal *et al.*². **c**, Evolution of the pressure of fluid saturation with f_{O_2} of a basalt melt having 40, 132 and 240 p.p.m. CO_2 and 0.1 wt% H_2O , which are minimum, average and maximum CO_2 contents, respectively, of the Siqueiros melt inclusions². At an f_{O_2} below $\Delta NNO = -1$, the pressure of saturation in fluid rises because of the continuous increase in CO of the coexisting gas phase. Grey horizontal line corresponds to the average collection pressure of Siqueiros basalts.

MgO of MORB documented worldwide⁶, the Siqueiros magmas would be expected near the lower end of the range (that is, $\Delta NNO = -3.5$; ref. 6). The CO_2 contents of Siqueiros melt inclusions average at 132 ± 34 p.p.m. but range from 43 p.p.m. up to 243 p.p.m. (ref. 2).

Figure 1c shows the evolution of the pressure of fluid saturation with f_{O_2} of basalt melts having 40, 132 and 240 p.p.m. CO_2 and 0.1 wt% H_2O . It can be seen that, except for the lowest CO_2 contents, most melts would be fluid-saturated at their collection pressure for an $f_{O_2} < \Delta NNO = -2.5$. Considering the uncertainties associated with the determination of dissolved CO_2 in MORB glasses (± 15 p.p.m.) and with the redox state of Siqueiros magmas, we contend that the condition of fluid saturation before eruption cannot be disregarded for at least the most CO_2 -rich Siqueiros melt inclusions.

We note that this condition is in agreement with earlier findings showing that the redox state of oceanic basalts is compatible with mantle melting under fluid-present or graphite-saturated conditions^{7,8}. Therefore, although the variable CO_2 content of quenched oceanic basaltic glasses results from syneruptive degassing³, part of this variability may also reflect regional-to-local variations in f_{O_2} . In general, a quantitative modelling of volatiles' behaviour in MORB magmas will require explicit consideration of the role of f_{O_2} (ref. 9).

Bruno Scaillet*, Michel Pichavant

*ISTO-CNRS, UMR 6113, 45071 Orléans cedex 2, France

e-mail: bscaille@cnrs-orleans.fr

doi:10.1038/nature02814

- Wallace, P., Anderson, A. T. & Davis, A. M. *Nature* **377**, 612–616 (1995).
- Saal, A. E., Hauri, E. H., Langmuir, C. H. & Perfit, M. R. *Nature* **419**, 451–455 (2002).
- Dixon, J., Stolper, E. M. & Holloway, J. R. *J. Petrol.* **36**, 1607–1631 (1995).
- Holloway, J. R. in *Reviews in Mineralogy* (eds Carmichael, I. S. E. & Eugster, H. P.) **17**, 211–233 (Mineralogical Society of America, Book Crafters, Chelsea, Michigan, 1987).
- Pawley, A. R., Holloway, J. R. & McMillan, P. F. *Earth Planet. Sci. Lett.* **110**, 213–225 (1992).
- Christie, D. M., Carmichael, I. S. E. & Langmuir, C. H. *Earth Planet. Sci. Lett.* **79**, 397–411 (1986).
- Blundy, J. D., Brodholt, J. P. & Wood, B. J. *Nature* **349**, 321–324 (1991).
- Holloway, J. R. *Chem. Geol.* **147**, 89–97 (1998).
- Mathez, E. *Nature* **310**, 371–375 (1984).

Saal et al. reply — Scaillet and Pichavant¹ raise an important point about the role that f_{O_2} plays in determining C–H–O fluid speciation and in estimating the degree of vapour saturation in oceanic basalts. However, this does not seem to be relevant to the volatile geochemistry of mid-ocean-ridge basalt (MORB) magmas in general and of Siqueiros MORB in particular.

To address their comments, we should first mention that the compositions of the Siqueiros picritic glasses are representative

of MORB. Their primitive composition does not indicate an unusually large extent of depletion compared with normal MORB, and indicates that they were not affected by the pervasive fractionation, mixing and aggregation processes taking place at the centre of the ridge segments^{2–4}.

The key factor in Scaillet and Pichavant's comment is the fO_2 of MORB. The authors incorrectly suggest that the Saal *et al.*⁴ estimation of the fO_2 of Siqueiros lavas falls at the upper range of the fO_2 estimated for MORB. Saal *et al.*⁴ calculated the fO_2 for Siqueiros picritic glasses and melt inclusions as $\Delta NNO = -1.7 \pm 0.5$ (2σ), using the compositions of syngenetic chromium–spinel inclusions in olivine phenocrysts⁵. The fO_2 value for Siqueiros samples agrees very well with previous estimations of the fO_2 for MORB, ranging from $\Delta NNO = -2.10 \pm 95$ (2σ) (ref. 6) to $\Delta NNO = -1.32 \pm 0.86$ (2σ) (ref. 7).

Our estimate of the fO_2 for Siqueiros samples, $\Delta NNO = -1.75$, is therefore a conservative value. Under these conditions ($\Delta NNO \geq -2$), the amount of CO existing with CO_2 is negligible (Fig. 1c of ref. 1). Furthermore, Scaillet and Pichavant¹ estimate an fO_2 of $\Delta NNO = -3.5$ for the Siqueiros samples, invoking a global correlation between $Fe^{+3}/\Sigma Fe$ ratios and MgO content in MORB⁶. However, the existing data^{6,7} show no clear correlation, but variation in $Fe^{+3}/\Sigma Fe$ ratios from 0.02 to 0.06 at any given MgO content.

The conclusions of Scaillet and Pichavant¹ also depend critically on the fugacity of molecular hydrogen (fH_2) in MORB and their discussion relies on an estimate for a high fH_2 in MORB (1 bar at 400 bars total pressure). Their Fig. 1a shows that similar calculations made with a lower fH_2 result in

no production of CO, and that the carbon speciation is dominated by CO_2 . Measurements of the composition of fluids trapped in MORB vesicles show that reduced vapour species (CO , H_2 , CH_4) are typically less than 0.03–1 vol% of the total vapour, even at low water content^{8,9}. These measurements point to the very low abundance of reduced C–H species in the fluids that are in equilibrium with MORB, and indicate that the estimates of vapour-saturation pressure in Siqueiros MORB (and indeed, probably all MORB) are accurate.

Several other observations have not been addressed by Scaillet and Pichavant. First, their estimated vapour-saturation pressures at $fO_2 \Delta NNO < -2$ (Fig. 1c of ref. 1) assume a finite solubility for CO in basaltic melt, but in fact the solubility of CO in basalt has not been directly measured. As a result, even if a basaltic melt did have a significant CO content, it is not yet possible accurately to estimate the vapour-saturation pressure of basalt in the presence of a mixed CO – CO_2 fluid^{10,11}.

Second, if vapour exsolution from the Siqueiros magmas had been important, we would expect that lavas that underwent degassing would contain a large amount of vesicles. However, the very low (0–0.5 vol%) vesicularity of the Siqueiros host glasses is consistent with their range from nearly saturated to undersaturated in H_2O – CO_2 vapour at the pressure of eruption⁴.

Third, significant degassing of CO would have destroyed the observed correlation between CO_2 , Nb and Cl contents in the Siqueiros samples⁴. CO_2 –Nb–Cl correlations are very difficult to explain if degassing of CO was important in those samples. Furthermore, Siqueiros melt inclusions and host glasses have the highest CO_2 /Nb ratios

of all MORB analysed, even though MORB is generally supersaturated in H_2O – CO_2 at the depth of eruption (because the rate of cooling is faster than that of bubble nucleation)^{4,12}. The simplest explanation for the CO_2 –Nb–Cl correlation and the high CO_2 /Nb ratios for the Siqueiros samples is that the picritic glasses and inclusions are undegassed.

Finally, it is clear that in the case of the Siqueiros melt inclusions they were trapped at pressures higher than the pressure of eruption of the lavas. Therefore the 400-bars pressure used by Saal *et al.*⁴ is a minimum estimation of the pressure of entrapment. If the inclusions were trapped at a pressure higher than 400 bars, their dissolved volatile contents would be lower than those required to saturate a basaltic melt in H_2O – CO_2 vapour.

Alberto E. Saal*, Erik H. Hauri, Charles H. Langmuir, Michael R. Perfit

**Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA
e-mail: asaal@brown.edu
doi:10.1038/nature02815*

1. Scaillet, B. & Pichavant, M. *Nature* doi:10.1038/nature02814 (2004).
2. Perfit, M. R. *et al.* *Earth Planet. Sci. Lett.* **141**, 91–108 (1996).
3. Sims, K. W. *et al.* *Geochem. Cosmochim. Acta* **66**, 3481–3504 (2002).
4. Saal, A. E., Hauri, E. H., Langmuir, C. H. & Perfit, M. R. *Nature* **419**, 451–455 (2002).
5. Danyushevsky, L. & Sobolev, A. V. *Mineral. Petrol.* **57**, 229–241 (1996).
6. Christie, D. M., Carmichael, I. S. E. & Langmuir, C. H. *Earth Planet. Sci. Lett.* **79**, 397–411 (1986).
7. Bezos, A. & Humler, E. *Geochem. Cosmochim. Acta* (in the press).
8. Javoy, M. & Pineau, F. *Earth Planet. Sci. Lett.* **107**, 598–611 (1991).
9. Cartigny, P., Jendrzejewski, N., Pineau, F., Petit, E. & Javoy, M. *Earth Planet. Sci. Lett.* **194**, 241–257 (2001).
10. Pawley, A. R., Holloway, J. R. & McMillan, P. F. *Earth Planet. Sci. Lett.* **110**, 213–225 (1992).
11. Holloway, J. R. *Chem. Geol.* **147**, 89–97 (1998).
12. Dixon, J. E. & Stolper, E. M. *J. Petrol.* **36**, 1633–1646 (1995).

The nonlinear nature of friction

Michael Urbakh¹, Joseph Klafter¹, Delphine Gourdon² & Jacob Israelachvili²

¹School of Chemistry, Raymond and Beverley Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel

²Department of Chemical Engineering, University of California at Santa Barbara, Santa Barbara, California 93106, USA

Tribology is the study of adhesion, friction, lubrication and wear of surfaces in relative motion. It remains as important today as it was in ancient times, arising in the fields of physics, chemistry, geology, biology and engineering. The more we learn about tribology the more complex it appears. Nevertheless, recent experiments coupled to theoretical modelling have made great advances in unifying apparently diverse phenomena and revealed many subtle and often non-intuitive aspects of matter in motion, which stem from the nonlinear nature of the problem.

Fric tion plays a central role in diverse systems and phenomena that at first sight may seem unrelated, but which on closer scrutiny are found to display common features that are shared by all tribological processes, in technological, geological or biological areas. The development of durable and/or low-friction surfaces and thin lubricating films has become an important factor in the miniaturization of moving components in many technological devices. These include micro-electro-mechanical systems (MEMS), computer recording systems and miniature motors with small loads. The old, simple empirical laws of friction do not always hold in such systems; this is due to their high surface-to-volume ratio and the greater importance of surface chemistry, adhesion and surface structure or roughness. Conventional tribological and lubrication techniques used for large objects can be ineffective at the nanometre scale, which requires new methods for control. Another rapidly growing area of tribology is in biosystems, and particularly the lubrication mechanisms in joints. Through the process of natural selection, nature has produced water-based lubricant systems that far outclass the best oil-based lubricants of most man-made devices; to emulate these systems is one of today's great challenges.

At the conceptual and theoretical levels, however, recent advances have revealed the enormous complexity of even the simplest tribological process. Friction is intimately related to both adhesion and wear, and all three require an understanding of highly non-equilibrium processes occurring at the molecular level to determine what happens at the macroscopic level. Surfaces can be smooth or rough, hard or soft, elastic, viscoelastic or plastic, brittle or ductile, dry (unlubricated) or lubricated, and of very different chemistries. The multitude of asperities on two shearing surfaces are constantly coming into and out of contact, where the local pressure between them can fluctuate between ~ 1 Pa (10^{-5} atmospheres of pressure) and GPa (10^4 atmospheres) within microseconds. These are extreme conditions that cannot always be treated by simple 'linear' theories.

Modern views of friction

To understand the behaviour of two real surfaces in relative motion while still in contact, we need to look into what is going on at the 'single asperity' level. With the advent of the atomic force microscope¹ (AFM) and the surface forces apparatus² (SFA) it became possible to study individual sliding junctions at the molecular level. The AFM and SFA are ideal tools in nano-, micro- and macroscopic tribological experiments for measuring the normal and lateral forces, and wear, between (1) a nanometre-radius tip or micro-metre-sized colloidal particle against a substrate surface, and (2) with the SFA, two macroscopic molecularly smooth or rough surfaces of measurable molecular contact area that confine a lubricant film of measurable thickness.

Three theoretical approaches that have been introduced to

investigate frictional forces in sheared systems are illustrated in Fig. 1: large-scale molecular dynamics (MD) simulations^{3–9}, phenomenological rate-state (RS) models^{10–15}, and 'minimalistic' models (MM)^{16–18}. Each approach has its advantages and disadvantages, and a different emphasis.

A tribological model is expected to recover certain key experimental observations, some of which are shown in the left panels of Fig. 2, as (1) structural transitions in thin lubricating liquid films induced by confining surfaces^{19,20} and how these are related to (2) periodic and chaotic stick-slip motion^{21–23} (Fig. 2a, b), where the shape of the stick-slip can be sawtooth or oscillatory²⁴, (3) transitions between 'smooth' and stick-slip sliding at certain critical sliding velocities or loads^{24,25} (Fig. 2c), (4) the very high effective viscosities of confined liquid films²⁶, and (5) the funicity, that is, dependence on the previous history, of friction forces^{21,27}.

Atomistic MD simulations (Fig. 1) have a wide range of applicability and have reached a high level of rigour and accuracy. They help us to understand liquid layering in nano-confinement²⁸, the relationship between static and kinetic friction^{6,10}, the nature of transitions between stick-slip and smooth sliding⁴, slippage at solid-liquid interfaces^{4,29,30}, shear thinning²⁹ and the friction of rough surfaces⁹. But MD simulations are currently limited to timescales no greater than tens of nanoseconds and length scales of tens of nanometres, which are too short for analysing many tribological systems⁷.

An important issue, therefore, was how to reduce the large-scale, many-parameter MD simulations to simpler descriptions with only a few equations of motion. Various phenomenological RS models^{10–15} provided such a description (Fig. 1), where the coefficients of one or two dynamical equations are fitted to experiment variables and then used to describe a wide range of observed frictional behaviours, such as the dilation of a liquid under shear¹⁵ and the transition between stick-slip (regular or chaotic) and smooth sliding friction^{12,31}. However, most 'state variables' in RS models cannot yet be quantitatively related to physical system properties^{11,12}.

An additional understanding of friction came with the MM (Fig. 1) that focuses on a small number of the most relevant degrees of freedom of confined molecules but can nevertheless explain phenomena of high complexity^{16–18,32,33}. Moreover, the MM enabled predictions to be made that were later verified experimentally^{23,25}. The MM naturally led to two characteristic states of the embedded system when sheared in the presence of thermal noise: 'trapped' and 'sliding' states. These are the ingredients that lead to stick-slip and the transition to sliding and are therefore the essential requirements for successful modelling of friction³⁴. The MM emphasized the nonlinear nature of frictional dynamics that has led to a potentially new method for controlling friction and/or boundary slip via the external manipulations.

Control of friction

The ability to control and manipulate frictional forces is extremely important for many applications. One may wish to reduce or enhance friction, modify the chaotic regime, and so on. Such control can be technologically important for micromechanical devices and computer disk drives, where the early stages of motion and the stopping processes often exhibit unwanted stick-slip or damage³⁵. In contrast, chaotic stick-slip may be desirable, for example, in string instruments. The control of frictional forces has been traditionally approached by chemical means, usually by supplementing base lubricants with friction modifier additives.

A completely different approach for 'tuning' frictional response, which has attracted considerable interest recently^{8,36–40}, is to control the system mechanically via normal vibrations of small amplitude and energy (Fig. 3). In this case, the idea is to reduce the friction force or to eliminate stick-slip motion through a stabilization of desirable modes of motion. Figure 3a and b show some recent experimental results^{36,37}, and Fig. 3c and d show corresponding theoretical MM³⁸ and MD⁸ modelling of these systems. Calculations demonstrated that oscillations of the normal load could lead to a transition from a state of high-friction stick-slip dynamics to a low-friction smooth sliding state. Manipulation by mechanical excitations, when applied at the right frequency, amplitude and direction, pull the molecules out of their potential energy minima

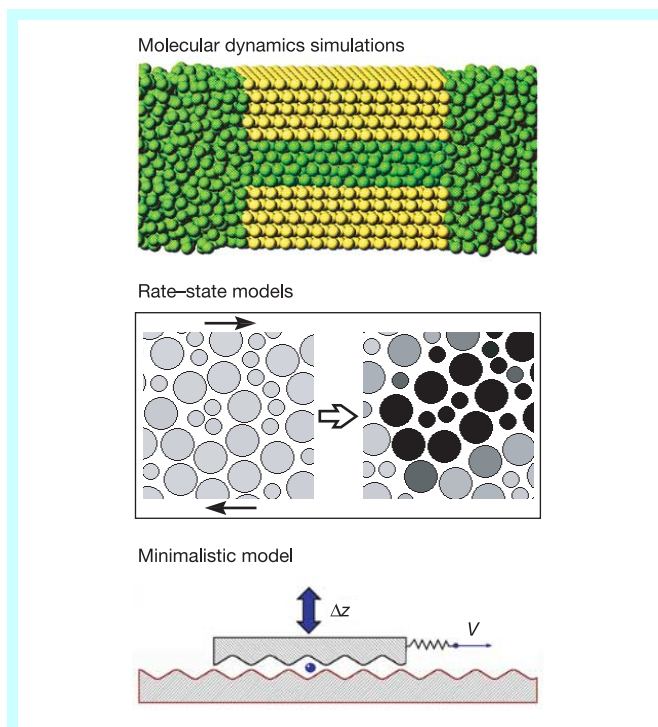


Figure 1 Three theoretical approaches to model friction. These are molecular dynamics simulations^{3–9}, rate-state models^{10–15}, and a 'minimalistic' model^{16–17,32,33}. MD simulations⁸, here showing shear-ordered lubricant molecules (green) between gold surfaces, follow the trajectory of molecules in space and time by solving the equations of motion as determined by the interatomic potential functions. RS models assume that the system is composed of 'local phases' (domains or grains) that can be described by a small number of 'state variables' that characterize the deformations, molecular rearrangements^{10,12}, dilation¹⁵ and other (statistical) properties of the interfacial material. Darker grains indicate a larger shear displacement with time. MM reduces the system to its bare essentials, for example, representing an embedded system by a single particle (which can also describe a system of many non-interacting particles) between two surfaces. Most of the experimental observations are qualitatively recovered by the MM.

and thereby reduce friction (at other frequencies or amplitudes the friction can be increased).

Friction and lubrication in biology

Animals, insects, their internal organs, tissues and biological microstructures and microorganisms experience much the same friction and lubrication forces in their movement as do machines. These can involve both lubricated (Fig. 4a) and unlubricated surfaces (Fig. 4b). The main difference between man-made and natural (biological) lubricants is that the former are usually 'oil-based' while the latter are 'water-based'. Water-based lubricant systems function well with hydrophilic surfaces where the surface charge provides an electrostatic 'double-layer' repulsion, in addition to the 'steric' repulsion of the hydration layer of tightly bound water molecules⁴¹. Recent experiments⁴² show that brushes of charged polymers (polyelectrolytes) attached to surfaces rubbing across an aqueous medium result in superior lubrication even at low sliding velocities and at pressures up to several atmospheres. Some biolubricating systems, such as in the eyes, may be similarly mediated by brush-like polyelectrolyte layers.

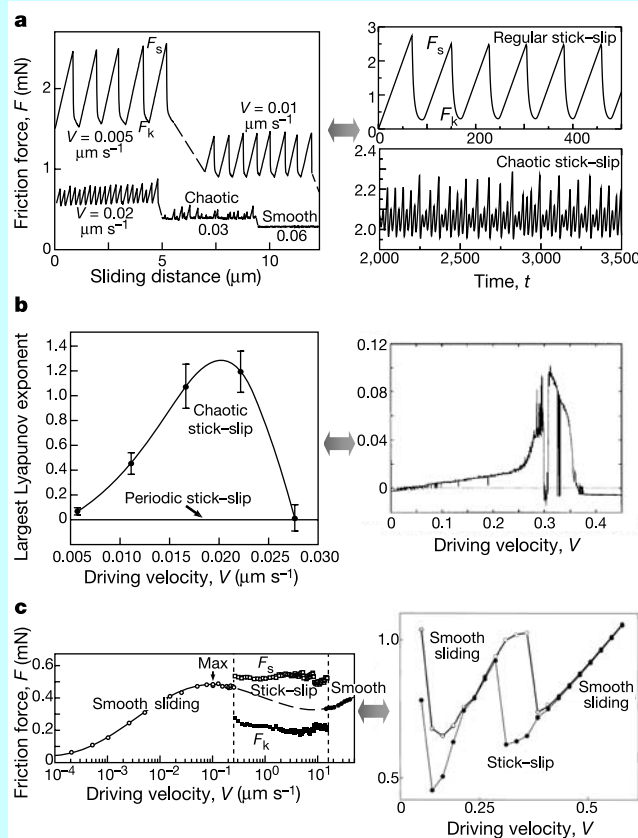


Figure 2 Examples of complex tribological effects of friction forces. Such effects versus sliding distance or time are shown as measured (left panels) and modelled (right panels) in dimensionless units. **a**, Friction traces of ultra-thin films of the model lubrication-oil squalane ($C_{30}H_{62}$) between two shearing mica surfaces showing a typical transition from periodic stick-slip to smooth sliding via a chaotic stick-slip regime, as measured²³ and modelled by the RS model³¹. F_s and F_k are the static and kinetic friction forces in the stick-slip regime. **b**, Positive Lyapunov exponent—an indicator of 'chaos' as opposed to random or 'stochastic' motion—as measured²³ and modelled by the MM model^{16,17}. **c**, Example of measured²⁵ and MM modelling³² of a transition from smooth to (inverted) stick-slip sliding and to smooth sliding again with increasing velocity. Open and closed symbols show the maximum and minimum friction forces in the inverted stick-slip regime. An RS model has also successfully explained this phenomenon²⁵.

Water-based lubricants are more efficient but cannot function at high temperatures owing to the high volatility and oxidative reactivity of water. But with the development of MEMS devices that are often made of ceramic materials and designed to operate at ambient temperatures, it is likely that we will learn from biology how to use water-based fluid lubricants. Another important differ-

ence between biological and man-made lubrication systems is that in the former the lubricant is often chemically attached to the surface, as occurs at the cartilage surfaces of joints. The recent development of computer disk surfaces, using 2–4-nm thick perfluoroether polymer layers chemically grafted to the carbon surfaces, is an example of this trend.

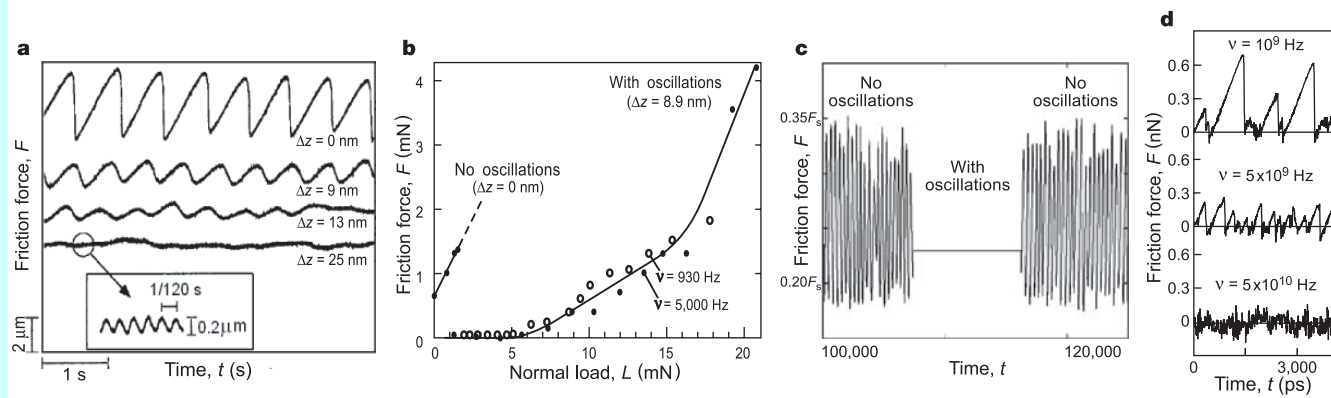


Figure 3 Reduction of friction and stick-slip by mechanical excitations. **a, b**, Experiments^{36,37} showing a drastic reduction in the stick-slip amplitude ($F_s - F_k$), the friction force F , or the friction coefficient $\mu = F/L$, by applying low-energy oscillations of frequency ν and small amplitude Δz perpendicular to the sliding direction. **c**, MM modelling³⁸ shows the elimination of stick-slip during the oscillations.

The friction force is given in units of static friction; time in dimensionless units. **d**, An MD simulation of F versus time at three different frequencies for a system of two organic surfactant-coated surfaces as in **b** showing the reduction of stick-slip and F with increasing ν .

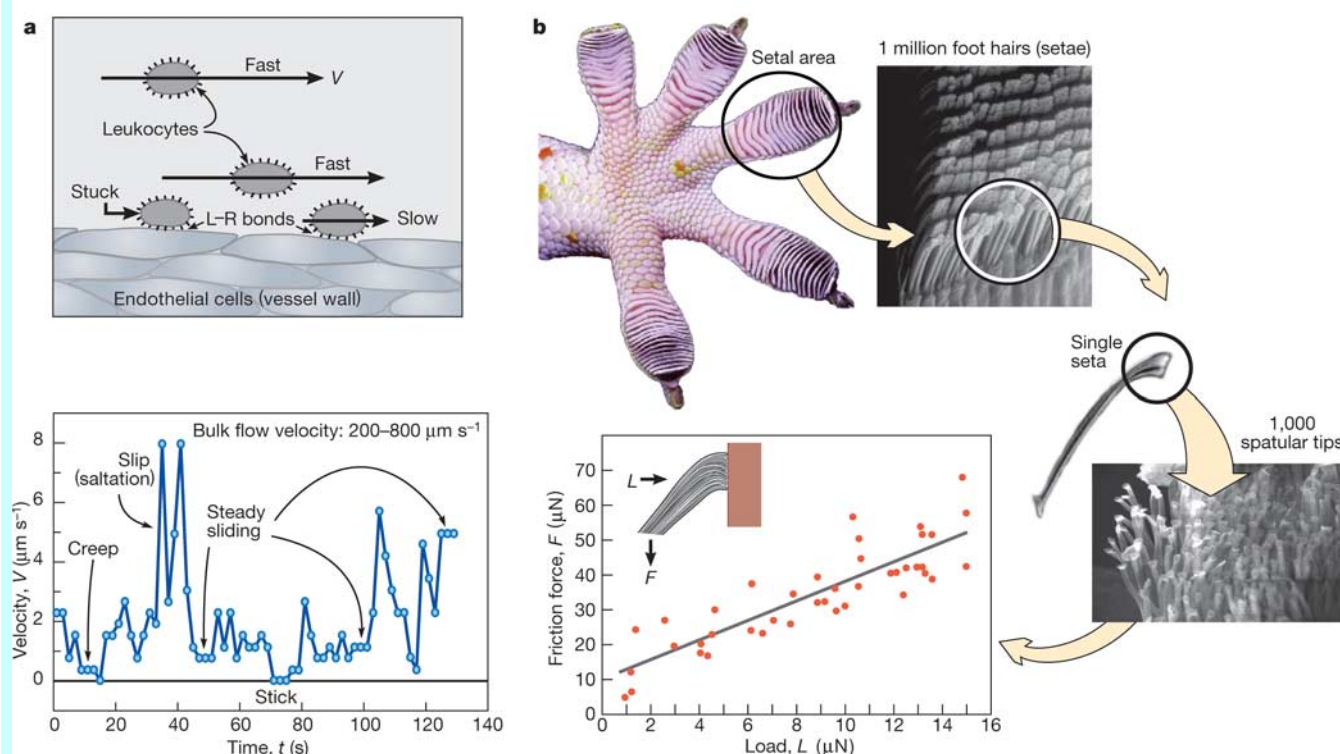


Figure 4 Two examples of friction and lubrication forces in living systems. **a**, *In vitro* 'rolling' of leukocyte cells while in contact with the endothelium (top) resembles chaotic stick-slip behaviour (bottom⁴²) involving sticking ($V = 0$), slow sliding (referred to as 'rolling' or creep), steady sliding, slip or 'saltation', and free or bulk flow ($V = 200$ –

$800 \mu\text{m s}^{-1}$). **b**, The high adhesion and friction forces between the pads on gecko feet and surfaces allow them to climb up walls. The truly amazing aspect of gecko locomotion is their ability to both attach and detach, that is, control high binding and unbinding forces, within a few milliseconds. Figures kindly supplied by K. Autumn.

Theoretically, the differences between water-based and oil-based systems are probably more quantitative than qualitative, and there is no reason to believe that existing nonlinear dynamical models which apply to machines and nano-devices are not also applicable to biological systems, such as those shown in Fig. 4. Thus, the saltatory motion of white blood cells along the endothelium surfaces of blood capillaries (Fig. 4a, top) follows a chaotic-like stick-slip motion (Fig. 4a, bottom). This behaviour has been successfully modelled in terms of the thermally controlled binding and unbinding of specific ligand-receptor bonds on the two surfaces^{43,44}.

Figure 4 shows examples of how adhesion and friction are intimately coupled in complex macromolecular systems. This coupling is also found in other biological systems, such as protein unfolding, which often follows a stick-slip process⁴⁵. The close relationship between adhesion, stick-slip and friction, which ultimately involve the making and breaking of bonds, is at the heart of recent theoretical and experimental studies^{46–50} that suggested a universal $|\ln V|^{2/3}$ - dependence of the height of stick-slip spikes on the driving velocity V .

Outstanding fundamental questions

What one hopes for is a unified approach to energy-dissipating systems that encompasses most tribological but also other phenomena, for example, in biology and geology. The models should be able to explain complex but common observations in terms of meaningful physical quantities and unravel the origin of energy dissipation which underlies all friction processes. Later, one would like to use these models for making predictions. More specifically, some of the important questions are:

- (1) Why is 'static friction' so universally observed between solid objects?
- (2) How are friction and wear related? And why does surface damage often occur at the start of motion?
- (3) How are the static and kinetic friction forces, and the characteristic transition velocities between smooth and stick-slip sliding, determined by the molecule-molecule and molecule-surface interactions and, in macroscopic systems, asperity-asperity or grain-grain interactions?
- (4) Are the stick and slip regimes indicative of different phase states (liquid, solid, glassy) of the confined films or interfaces?
- (5) What 'hidden' information is contained in chaotic as opposed to periodic motion (compare Fig. 2)? This is particularly important for predicting earthquakes.
- (6) And finally, how can we control friction in practice, most often to reduce it or eliminate stick-slip at all pressures and velocities? But there are also situations when one wants high friction, as in clutches and brakes, or stick-slip, to enrich the sound of a violin and improve the feel or 'texture' of processed food as sensed during biting and chewing. □

Received 24 October 2003; accepted 9 June 2004; doi:10.1038/nature02750.

1. Binnig, G., Quate, C. F. & Gerber, Ch. The atomic force microscope. *Phys. Rev. Lett.* **56**, 930–933 (1996).
2. Israelachvili, J. N. & Adams, G. E. Measurement of forces between two mica surfaces in aqueous electrolyte solutions in the range 0–100 nm. *J. Chem. Soc. Faraday Trans. 1* **74**, 975–1001 (1978).
3. Landman, U., Luedtke, W. D. & Ringer, E. M. in *Fundamentals of Friction: Macroscopic and Microscopic Processes* (eds Singer, I. L. & Pollock, H. M.) 463–510 (Kluwer, Dordrecht, 1992).
4. Thompson, P. A. & Robbins, M. O. Origin of stick-slip motion in boundary lubrication. *Science* **250**, 792–794 (1990).
5. Muser, M. H., Urbakh, M. & Robbins, M. O. Statistical mechanics of static and low-velocity kinetic friction. *Adv. Chem. Phys.* **126**, 187–272 (2003).
6. He, G., Muser, M. H. & Robbins, M. O. Adsorbed layers and the origin of static friction. *Science* **284**, 1650–1652 (1999).
7. Robbins, M. O. & Muser, M. H. in *Modern Tribology Handbook* (ed. Bhushan, B.) 717–757 (CRC Press, Boca Raton, Florida, 2001).
8. Gao, J. P., Luedtke, W. D. & Landman, U. Friction control in thin film lubrication. *J. Phys. Chem. B* **102**, 5033–5037 (1998).
9. Gao, J. P. *et al.* Frictional forces and Amontons' Law: from the molecular to the macroscopic scale. *J. Phys. Chem. B* **108**, 3410–3425 (2004).
10. Persson, B. N. J. *Sliding Friction, Physical Properties and Applications* (Springer, Berlin, 2000).

11. Ruina, A. Slip instability and state variable friction laws. *J. Geophys. Res.* **88**, 10359–10370 (1983).
12. Carlson, J. M. & Batista, A. A. Constitutive relation for the friction between lubricated surfaces. *Phys. Rev. E* **53**, 4153–4165 (1996).
13. Urbakh, M., Daikhin, L. & Klafter, J. Dynamics of confined liquids under shear. *Phys. Rev. E* **51**, 2137–2141 (1995).
14. Aranson, I. S., Tsirring, L. S. & Vinokur, V. M. Stick-slip friction and nucleation dynamics of ultrathin liquid films. *Phys. Rev. B* **65**, 125402 (2002).
15. Lemaître, A. Rearrangements and dilatancy for sheared dense materials. *Phys. Rev. Lett.* **89**, 195503 (2002).
16. Rozman, M. G., Urbakh, M. & Klafter, J. Stick-slip motion and force fluctuations in a driven two-wave potential. *Phys. Rev. Lett.* **77**, 683–686 (1996).
17. Rozman, M. G., Urbakh, M. & Klafter, J. Origin of stick-slip motion in a driven two-wave potential. *Phys. Rev. E* **54**, 6485–6494 (1996).
18. Muser, M. H., Wenning, L. & Robbins, M. O. Simple microscopic theory of Amontons's laws for static friction. *Phys. Rev. Lett.* **86**, 1295–1298 (2001).
19. Drummond, C., Alcantar, N. A. & Israelachvili, J. N. Shear alignment of confined hydrocarbon liquid films. *Phys. Rev. E* **66**, 011705 (2002).
20. Klein, J. & Kumacheva, E. Confinement-induced phase-transitions in simple liquids. *Science* **269**, 816–819 (1995).
21. Demirel, A. L. & Granick, S. Friction fluctuations and friction memory in stick-slip motion. *Phys. Rev. Lett.* **77**, 4330–4333 (1996).
22. Drummond, C. & Israelachvili, J. N. Dynamic behavior of confined branched hydrocarbon lubricant fluids under shear. *Macromolecules* **33**, 4910–4920 (2000).
23. Drummond, C. & Israelachvili, J. N. Dynamic phase transitions in confined lubricant fluids under shear. *Phys. Rev. E* **63**, 041506 (2001).
24. Gourdon, D. & Israelachvili, J. Transitions between smooth and complex stick-slip sliding of surfaces. *Phys. Rev. E* **68**, 021602 (2003).
25. Drummond, C., Israelachvili, J. & Richetti, P. Friction between two weakly adhering boundary lubricated surfaces in water. *Phys. Rev. E* **67**, 066110 (2003).
26. Hu, H. W., Carson, G. A. & Granick, S. Relaxation-time of confined liquids under shear. *Phys. Rev. Lett.* **66**, 2758–2761 (1991).
27. Yoshisawa, H., Chen, Y.-L. & Israelachvili, J. Fundamental mechanisms of interfacial friction I: Relation between adhesion and friction. *J. Phys. Chem.* **97**, 4128–4140 (1993).
28. Gao, J. P., Luedtke, W. D. & Landman, U. Layering transitions and dynamics of confined liquid films. *Phys. Rev. Lett.* **79**, 705–708 (1997).
29. Thompson, P. A., Robbins, M. O. & Grest, G. S. Structure and shear response in nanometer-thick films. *Isr. J. Chem.* **35**, 93–106 (1995).
30. Barrat, J.-L. & Boquet, L. Influence of wetting properties on hydrodynamic boundary conditions at a fluid/solid interface. *Faraday Discuss.* **112**, 1–9 (1999).
31. Lemaître, A. & Carlson, J. Boundary lubrication with a glassy interface. *Phys. Rev. E* (in the press).
32. Filippov, A. E., Klafter, J. & Urbakh, M. Inverted stick-slip friction: what is the mechanism? *J. Chem. Phys.* **116**, 6871–6874 (2002).
33. Rozman, M. G., Urbakh, M., Klafter, J. & Elmer, F.-J. Atomic scale friction and different phases of motion of embedded molecular systems. *J. Phys. Chem. B* **102**, 7924–7930 (1998).
34. Filippov, A. E., Klafter, J. & Urbakh, M. Confined molecules under shear: from a microscopic description to phenomenology. *Phys. Rev. Lett.* **87**, 275506 (2001).
35. Bhushan, B. (ed.) *Micro/Nanotribology and Its Applications Series E Applied Sciences* Vol. 330, 1–668 (NATO Advanced Sciences Institutes, Kluwer Academic, Dordrecht/Boston/London, 1997).
36. Cochard, A., Bureau, L. & Baumberger, T. Stabilization of frictional sliding by normal load modulation. *Trans. ASME* **70**, 220–226 (2003).
37. Heuberger, M., Drummond, C. & Israelachvili, J. N. Coupling of normal and transverse motions during frictional sliding. *J. Phys. Chem. B* **102**, 5038–5041 (1998).
38. Rozman, M. G., Urbakh, M. & Klafter, J. Controlling chaotic friction. *Phys. Rev. E* **57**, 7340–7343 (1998).
39. Zaloj, V., Urbakh, M. & Klafter, J. Modifying friction by manipulating normal response to lateral motion. *Phys. Rev. Lett.* **82**, 4823–4826 (1999).
40. Braiman, Y., Barhen, J. & Protopopescu, V. Control of friction at the nanoscale. *Phys. Rev. Lett.* **90**, 094301 (2003).
41. Israelachvili, J. N. Measurement of the viscosity of liquids in very thin films. *J. Colloid Interf. Sci.* **110**, 263–271 (1986).
42. Raviv, U. *et al.* Lubrication by charged polymers. *Nature* **425**, 163–165 (2003).
43. Chang, K.-C. & Hammer, D. A. Adhesive dynamics simulations of Sialyl^x-Lewis^x/E-selectin-mediated rolling in a cell-free system. *Biophys. J.* **79**, 1891–1902 (2000).
44. Goetz, D. J., El-Sabban, M. E., Pauli, B. U. & Hammer, D. A. Dynamics of neutrophil rolling over stimulated endothelium *in vitro*. *Biophys. J.* **66**, 2202–2209 (1994).
45. Evans, E. Probing the relation between force-lifetime-and chemistry in single molecular bonds. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 105–128 (2001).
46. Sang, Y., Dube, M. & Grant, M. Thermal effects on atomic friction. *Phys. Rev. Lett.* **87**, 174301 (2001).
47. Dudko, O., Filippov, A. E., Klafter, J. & Urbakh, M. Dynamical force spectroscopy: a Fokker-Planck approach. *Chem. Phys. Lett.* **352**, 499–504 (2002).
48. Dudko, O., Filippov, A. E., Klafter, J. & Urbakh, M. Beyond the conventional description of dynamic force spectroscopy of adhesion bonds. *Proc. Natl Acad. Sci. USA* **100**, 11378–11381 (2003).
49. Riedo, E., Gneco, E., Bennewitz, R., Meyer, E. & Brune, H. Interaction potential and hopping dynamics governing sliding friction. *Phys. Rev. Lett.* **91**, 084502 (2003).
50. Stills, S. & Overney, R. Creeping friction dynamics and molecular dissipation mechanisms in glassy polymers. *Phys. Rev. Lett.* **91**, 095501 (2003).

Crystal structure of the calcium pump with a bound ATP analogue

Chikashi Toyoshima & Tatsuki Mizutani*

Institute of Molecular and Cellular Biosciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-0032, Japan

* Present address: Department of Immunology, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan

P-type ATPases are ATP-powered ion pumps that establish ion concentration gradients across cell and organelle membranes. Here, we describe the crystal structure of the Ca^{2+} pump of skeletal muscle sarcoplasmic reticulum, a representative member of the P-type ATPase superfamily, with an ATP analogue, a Mg^{2+} and two Ca^{2+} ions in the respective binding sites. In this state, the ATP analogue reorganizes the three cytoplasmic domains (A, N and P), which are widely separated without nucleotide, by directly bridging the N and P domains. The structure of the P-domain itself is altered by the binding of the ATP analogue and Mg^{2+} . As a result, the A-domain is tilted so that one of the transmembrane helices moves to lock the cytoplasmic gate of the transmembrane Ca^{2+} -binding sites. This appears to be the mechanism for occluding the bound Ca^{2+} ions, before releasing them into the lumen of the sarcoplasmic reticulum.

P-type ATPases are ion pumps that transfer cations across lipid bilayers. They establish ion concentration gradients that are used in a variety of biological processes and are referred to as such because they form a key phosphorylated intermediate in the reaction cycle (Fig. 1, inset; see ref. 1 for a review). The pumping of ions is thought to be achieved by altering the affinity and accessibility of the transmembrane ion-binding sites. Classical E1/E2 theory postulates that, in the E1 state, the binding sites have high affinity and open to the cytoplasm, whereas in the E2 state, the binding sites have low affinity and face the extracellular or luminal side^{2–4}. One important

feature of the pumping process is that, before being released, bound cations are occluded; that is, become inaccessible from either side of the membrane. Binding in itself does not lock the cytoplasmic gate, and bound cations can be exchanged with those in the cytoplasm (reviewed in ref. 5). Occlusion is generally thought to require phosphoryl transfer from ATP to the ATPase. Release of cations into the extracellular or luminal medium takes place while the enzyme is phosphorylated. Such autophosphorylation of the enzyme is a unique feature of P-type ATPases, and is distinct from other ATPases such as the F_0F_1 type.

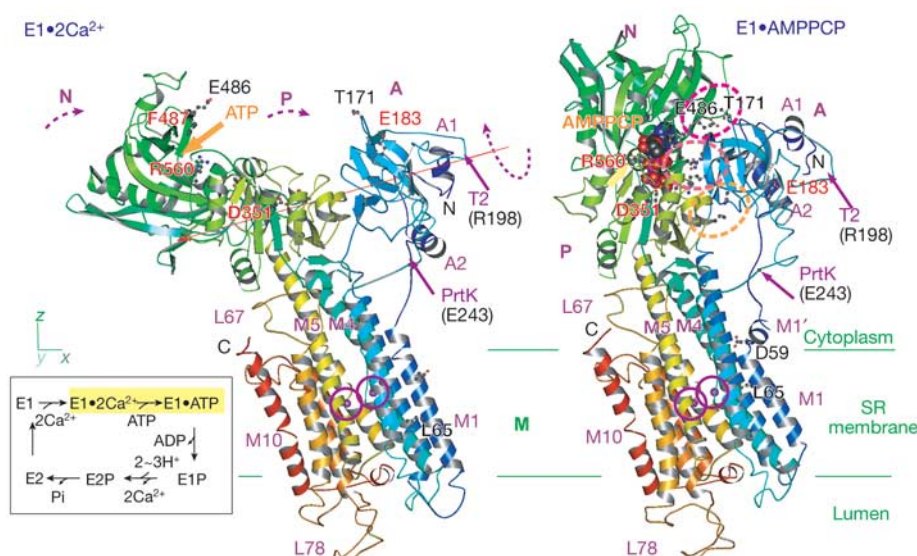


Figure 1 Front views (parallel to the membrane (x - y) plane) of Ca^{2+} -ATPase with (E1•AMPPCP) and without (E1• 2Ca^{2+}) AMPPCP (shown in space fill), an ATP analogue, in the presence of 10 mM Ca^{2+} . Colours change gradually from the amino terminus (blue) to the carboxy terminus (red). Purple spheres (circled) represent bound Ca^{2+} . Large arrows in E1• 2Ca^{2+} indicate the direction of movements of the cytoplasmic domains (A, N and P) in the E1• 2Ca^{2+} → E1•AMPPCP transition; the axis of tilting of the A-domain is also specified (thin red line). α -Helices in the A-domain (A1 and A2) and those in the transmembrane domain (M1, M4, M5 and M10) are indicated. M1' is an amphipathic part lying on the bilayer surface. PrtK, a proteinase-K digestion site (around Glu 243 (ref. 16));

SR, sarcoplasmic reticulum; T2, a trypsin digestion site at Arg 198 (ref. 10); ATP, the binding pocket for the adenosine moiety of ATP. Several key residues—E183 (A), F487 (N, adenine binding), D351 (P, phosphorylation site), R560 (N, β -phosphate binding) and those involved in interdomain hydrogen bonds (including T171 and E486)—are shown in ball-and-stick representation. Dotted circles indicate three contact spots between A- and N-domains (magenta) and between A- and P-domains (yellow). The figure was prepared with Molscript⁴⁴. Inset is a simplified reaction scheme (showing only the forward direction); the two states compared here are shown with a yellow background.

Of the P-type ATPase superfamily, the Ca^{2+} -ATPase (SERCA1a) of fast skeletal muscle sarcoplasmic reticulum is structurally^{6–8} and functionally⁹ the best-characterized member. It is an integral membrane protein with a relative molecular mass of 110,000 (M_r , 110K)^{10,11}, consisting of three (A (actuator), N (nucleotide binding) and P (phosphorylation)) cytoplasmic domains and ten (M1–M10) transmembrane helices⁶. Two high-affinity Ca^{2+} -binding sites are located side-by-side within the transmembrane region, formed by M4, M5, M6 and M8 helices⁶. We have already published crystal structures of the Ca^{2+} -ATPase in a Ca^{2+} -bound state⁶ (E1·2 Ca^{2+} ; Protein Data Bank (PDB) accession code 1SU4) and a Ca^{2+} -unbound state⁷ (E2(TG); PDB accession code 1IWO) stabilized by a potent inhibitor, thapsigargin (TG)¹². A brief overview of these structures is found in ref. 13.

Here we report the crystal structure of SERCA1a with a non-hydrolysable ATP analogue—adenosine 5′-[β , γ -methylene]triphosphate (AMPPCP)—and a Mg^{2+} bound to the cytoplasmic domains, and two Ca^{2+} ions at the transmembrane binding sites. This structure (which we abbreviate to E1·AMPPCP) probably represents the one just before phosphoryl transfer from ATP to the enzyme, and explains how the γ -phosphate and Mg^{2+} binding to the P-domain results in the occlusion of the bound Ca^{2+} some 50 Å away.

Structure determination

Two types of crystals belonging to different space groups ($P2_1$ and $C2$) were grown by dialysing affinity-purified enzyme against a buffer containing 0.5 mM AMPPCP in the presence of 10 mM Ca^{2+} and 5 mM Mg^{2+} . Diffraction data were highly anisotropic and showed good statistics to 2.9 Å resolution after merging. The structure was determined by generalized molecular replacement¹⁴, because heavy atom isomorphous replacement was unsuccessful.

The atomic model was refined with the $P2_1$ crystals to an R_{free} value of 29.6%.

Rearrangement of the cytoplasmic domains

The binding of AMPPCP and a metal ion to Ca^{2+} -ATPase in E1·2 Ca^{2+} causes large and global changes in the structure (Figs 1 and 2; see also Supplementary Movie), except for the transmembrane helices M6–M10. The most prominent difference is that the three cytoplasmic domains, widely separated in E1·2 Ca^{2+} , now form a compact headpiece. Gathering of the cytoplasmic domains also occurs in the E2(TG) state⁷ but the arrangements are different (Supplementary Fig. 1). Notably, the N-domain is more inclined towards the P-domain. This inclination is facilitated by AMPPCP, whose adenine ring binds to the N-domain around Phe 487 whereas the γ -phosphate binds to the P-domain around Asp 351, the residue of phosphorylation. As a result of the large inclination, the N-domain now makes contacts with the A-domain (Fig. 1) with an interface that is different compared with E2(TG). Despite the $\sim 90^\circ$ inclination, only small changes are observed within the N-domain itself (root mean square deviation (r.m.s.d.) = 0.74 Å) except for a flexible loop (Pro 500–Gly 509) at the top of the molecule. In contrast, the P-domain changes its structure (see below) and also its orientation with respect to the membrane plane ($\sim 15^\circ$; Fig. 1). As a result of these changes, the A-domain is tilted by $\sim 30^\circ$ around an axis approximately parallel to the membrane (specified in Fig. 1), bringing the M2 side of the A-domain higher up (Fig. 2a).

Locking of the cytoplasmic gate

The position of the A-domain in E1·AMPPCP is in marked contrast with that in E2(TG) (Supplementary Fig. 1), where the A-domain is rotated $\sim 110^\circ$ from the position in E1·2 Ca^{2+} around an axis

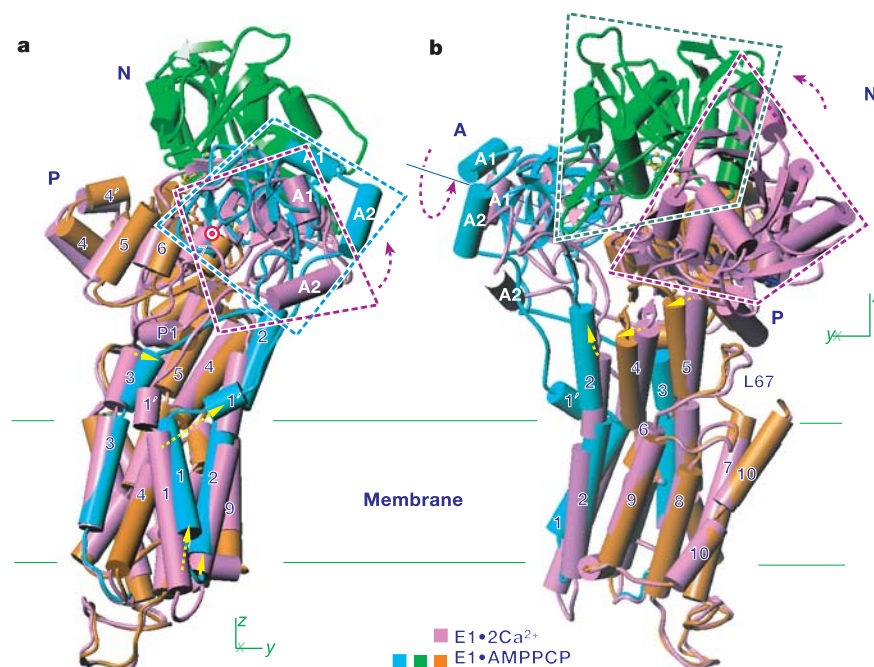


Figure 2 Superimposition of the E1·2 Ca^{2+} and E1·AMPPCP forms of Ca^{2+} -ATPase fitted with the transmembrane domain. E1·2 Ca^{2+} , violet; E1·AMPPCP, cyan (A-domain and M1–M3 helices), light green (N-domain) and orange (P-domain and M4–M10 helices). α -Helices are represented by cylinders and β -strands by arrows. Both (a and b) are viewed along the membrane plane, but at a difference of 45° . M2, M3 and M5 are represented by two or three cylinders, although they are continuous helices. Helices in the A-domain (A1 and A2), P-domain (4–7) and transmembrane helices are indicated. The

position of a preserved hydrogen bond between the A- and P-domains (Gly 156–Ala 725) is indicated by a double circle for E1·2 Ca^{2+} . Large arrows indicate the direction of movements of the cytoplasmic domains (A, N and P) in the E1·2 Ca^{2+} → E1·AMPPCP transition, and small ones those of transmembrane helices. Broken lines enclose the A-domain (a) and the N-domain (b). A thin blue line in b specifies the axis of tilting of the A-domain.

perpendicular to the membrane⁷. Because the A-domain is directly linked to the M1–M3 transmembrane helices, its movement inevitably causes rearrangements of these helices.

Of the three helices, M3 shows the smallest movement on nucleotide binding: only the top part is bent towards M2 by $\sim 20^\circ$ (Fig. 2a), although the loop connecting to the A-domain moves almost 20 Å (at Met 239) and appears to be strained, as indicated by the protection against proteinase K attack at Glu 243 (refs 15–17). The M2 helix shows a large and complex movement. M2 moves towards the cytoplasm by one turn of α -helix, presumably pulled by the tilt of the A-domain (Fig. 2a), and its cytoplasmic end shifts in a $+y$ -direction (Fig. 2b). The M1 helix shows a marked movement: it is pulled towards the cytoplasmic side by nearly two turns of α -helix ($\Delta z = 8.3$ Å for Leu 65; Fig. 1), and is bent largely at Asp 59, so that the amphipathic amino-terminal part (M1') lies on the membrane surface (Fig. 2a). This situation is very similar to that in E2(TG)⁷, and the part lying on the membrane surface (Trp 50–Glu 58) is identical. However, there is an $\sim 90^\circ$ difference in the direction of bending so that M1' approaches the M2 helix (that is, $+y$ - instead of $+x$ -direction as in E2(TG); Fig. 2a), reflecting the difference in orientation of the A-domain.

The functional meaning of the movement of the M1 helix is evident in the present structure. In E1·2Ca²⁺, which represents the state after the binding of both Ca²⁺ ions, Glu 309 caps the Ca²⁺ in site II (Fig. 3a, c)⁶. On its cytoplasmic side, there is a large empty space in which several water molecules were identified in the crystal structure (Fig. 3a)⁶. Thus, it seems possible for Glu 309 to adopt other side-chain conformations. In E1·AMPPCP, however, this space is occupied by M1 (Fig. 3b), and Leu 65 on the M1 helix makes van der Waals contacts with the Glu 309 side chain (Fig. 3b, c).

Glu 309 now forms a hydrogen bond with Asn 796 instead of with Glu 58 as in E1·2Ca²⁺ (Fig. 3). Thus, the conformation of Glu 309 is doubly locked.

It is well established that Ca²⁺ in site II is the second Ca²⁺ to bind^{18,19} and is exchangeable with Ca²⁺ in the cytoplasm^{5,20}. There is now strong evidence to indicate that Glu 309 works as the gating residue^{21,22}. It is easy to see that site II Ca²⁺ will be ready to dissociate, if the Glu 309 carboxyl detaches from it by thermal movement. Virtually no Ca²⁺ exchange takes place in E1·AlF_x·ADP, indicating that the Ca²⁺ ions are occluded. Moreover, the Glu309Gln mutant cannot occlude at all whereas the Asn796Ala mutant can to a certain extent²². On the other hand, biochemical data²⁰ show that binding of AMPPCP itself is not sufficient to occlude bound Ca²⁺, despite the fact that the crystal structures of E1·AMPPCP and E1·AlF_x·ADP (C.T., H. Nomura and T. Tsuda, unpublished data) are virtually the same except for details around the phosphorylation site. They also show very similar resistance to proteinase K attack at Glu 243 (Fig. 1)^{15,16}. Extra stability provided by AlF_x, a stable analogue of phosphate, may account for the difference in dissociation kinetics. Also, crystal packing may have selected a specific conformation from many possible ones in solution. Nonetheless, taken together, it seems clear that the locking of the conformation of Glu 309 by M1 helix is the mechanism of occlusion. A key residue here is Leu 65, as also demonstrated by a mutagenesis study²³.

In the E2(TG) form, the P-domain is inclined $\sim 30^\circ$ compared with the E1·2Ca²⁺ form⁷. This inclination brings about a large downward movement (~ 5.5 Å) of the M4 helix and destroys the Ca²⁺-binding sites. In E1·AMPPCP, the P-domain is inclined in the same direction but only by about half as much as in E2(TG), because

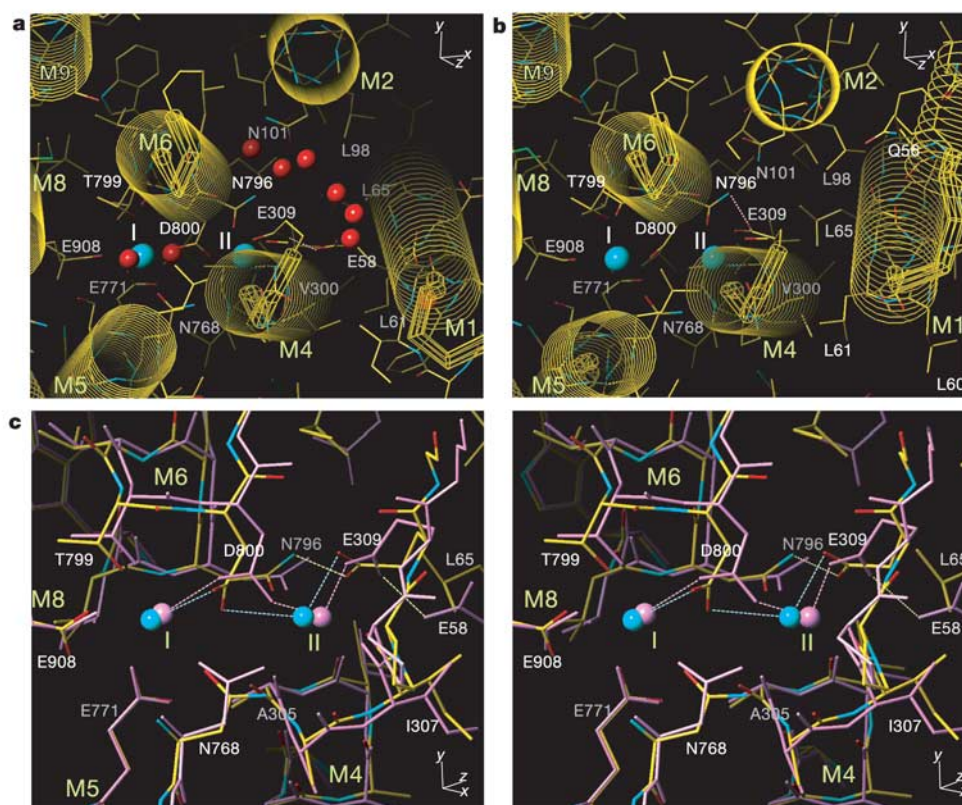


Figure 3 Transmembrane Ca²⁺-binding sites (I and II) and the movement of the M1 helix. **a**, E1·2Ca²⁺; **b**, E1·AMPPCP; **c**, superimposition of E1·2Ca²⁺ (violet) and E1·AMPPCP (atom colour) in stereo view. Cyan (**a–c**) and violet spheres (**c**) represent bound Ca²⁺; red

spheres indicate water molecules in the crystals. Owing to the resolution limitation of the diffraction data, no water molecules are shown in **b**. Dotted lines in **c** show the coordination of Ca²⁺ and hydrogen bonds involving E309 and D800 side chains.

bound Ca^{2+} restricts the bending of M5. The upper part of M4 certainly shows a downward movement (Fig. 2b)⁷. It is compensated, however, by the bending of the P-domain (see below) and is 'absorbed' by the unwound part containing Glu 309. To achieve this, Glu 309 appears to alter its main-chain conformation, but also pushes site II Ca^{2+} and Asp 800 (Fig. 3c). As a result, the coordination geometry of the Ca^{2+} becomes loose, particularly at site II. This change presumably prepares for the release of bound Ca^{2+} to the lumen during the conversion to E2P, the subsequent step in the reaction cycle.

ATP as a cleavable crosslinker

As described in the preceding sections, the large inclination of the N-domain is important for orientating the A-domain to close the cytoplasmic gate. AMPPCP facilitates this by crosslinking the N- and P-domains (Fig. 4). The conformation of AMPPCP appears to be rather uncommon: the phosphate moiety adopts a zigzag

configuration, similar to that in the histidine kinase CheA (PDB accession code 1I58)²⁴, bringing the β -phosphate close to the ribose. The orientation of the ribose and the configuration of the phosphates are different compared with those observed with the NMR structure of ATP bound to the isolated N-domain of Na^+K^+ -ATPase with no Mg^{2+} (ref. 25). The adenine ring of AMPPCP is positioned predominantly by stacking with Phe 487, consistent with previous results^{6,25,26} (Fig. 4a). Hydrogen bonding with main-chain carbonyl and amide groups, found with many adenosine-binding proteins²⁷, is not observed here. Instead, a ribose hydroxyl (O3^*) is stabilized by Arg 678, which in turn appears to form hydrogen bonds with residues both in the N- and P-domains (Fig. 4b). The O2^* of the ribose may also contribute to the positioning of the ribose by van der Waals contacts. These features are consistent with infrared spectroscopy studies²⁸. The α -phosphate is stabilized by Arg 489 (Fig. 4a) and the β -phosphate by Arg 560, which also appears to form hydrogen bonds with residues both in the N- and P-domains (Fig. 4b). Within a 3.4 Å distance from the γ -phosphate, there is a carboxyl group of Asp 351, hydroxyl groups of Thr 353 and Thr 625, amide groups of Thr 353 and Gly 626, a carbonyl group of Thr 353 and a metal ion (modelled as Mg^{2+} ; see Supplementary Methods) plus two coordinating water molecules. Asn 706 and Lys 684 are slightly more distant. Although the accuracy of such information is limited at 2.9 Å resolution, all of these residues have been identified as being sensitive to mutation^{29–31}. No bridging Mg^{2+} is found between the β - and γ -phosphates, presumably because Asp 627 blocks proper octahedral coordination of Mg^{2+} .

An intricate network of hydrogen bonds thus appears to be formed around AMPPCP (Fig. 4b), including those bridging the N- and P-domains. Of the residues involved in such bridging interactions, mutagenesis studies have shown that Arg 560 on the N-domain is particularly important^{17,32,33}. This residue appears to orientate the phosphate chain of AMPPCP in the right direction and form a salt bridge with Asp 627 in the P-domain, thereby establishing the N-domain–P-domain interaction. The importance of Arg 560 indicates that an oxygen atom of the β -phosphate has a key role in positioning this residue mobile in E1·2 Ca^{2+} and also suggests that the N-domain will move back when ADP leaves the enzyme. It is well known that acetylphosphate and carbamoylphosphate are substrates of Ca^{2+} -ATPase³⁴, and indeed both of them have an oxygen atom at this position.

Changes in the P-domain structure

On binding the γ -phosphate and Mg^{2+} , the P-domain changes its internal structure and its overall orientation with respect to the M5 helix (Fig. 5). As a member of the haloacid dehalogenase superfamily³⁵, the P-domain of Ca^{2+} -ATPase comprises a Rossmann fold that consists of a central seven-stranded (P β 1–P β 7) parallel β -sheet and associated helices (P α 1–P α 7)⁶. A unique feature of the β -sheet in Ca^{2+} -ATPase is that the two halves (P β 1–P β 4 and P β 5–P β 7) are staggered in unphosphorylated forms, but have a better alignment in phosphorylated (or γ -phosphate bound) forms (Fig. 5a). The top part of the first half (P β 1–P β 4) moves together as a result of binding of the γ -phosphate and Mg^{2+} , because Thr 353 positioned just above P β 1 coordinates to both ligands (Figs 4b and 5). Thr 625 and Gly 626 in the loop connecting P β 2 and P α 2 appear to coordinate to the γ -phosphate (Fig. 4). Furthermore, P β 5 twists upon binding of Mg^{2+} owing to the coordination by Asp 703 (Figs 4 and 5), which causes the tilting of P α 5–P α 7. Thus, the P-domain is bent in nearly two orthogonal directions (Fig. 5).

This change in the β -sheet structure brings the N-domain $\sim 30^\circ$ closer to the P-domain and is presumably essential for achieving a nearly 90° inclination required for ATP to reach the P-domain. Thus, the staggered β -sheet works as a secondary hinge, in addition to the main one (around Pro 602 and Pro 603) that covers $\sim 60^\circ$ as in the E1·2 $\text{Ca}^{2+} \rightarrow \text{E2(TG)}$ transition. This secondary hinge also changes the direction of movement of the N-domain (Fig. 5),

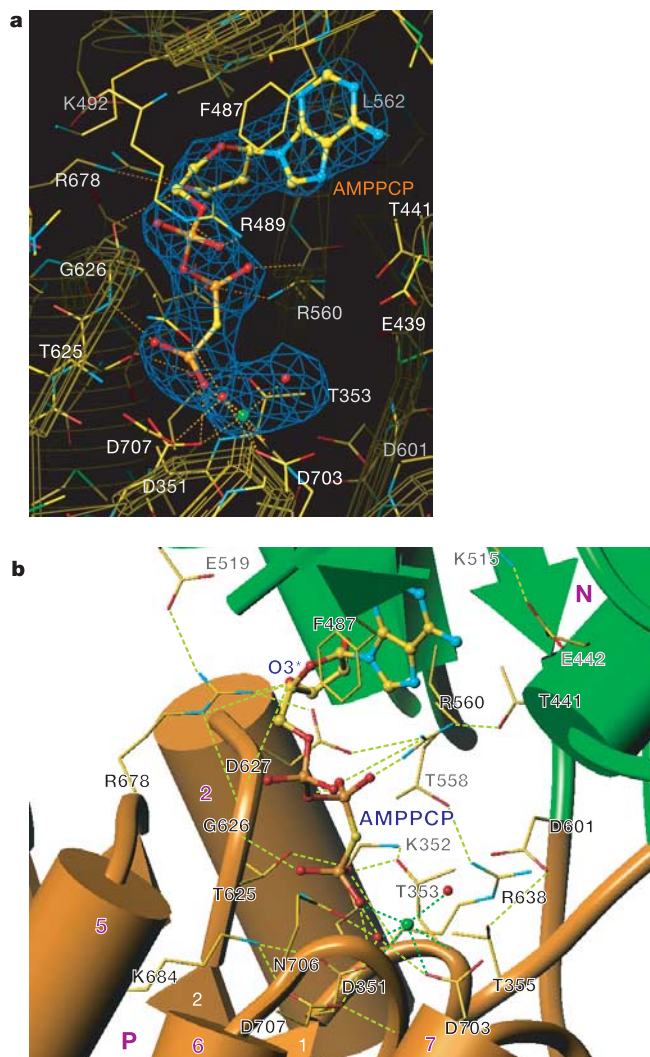


Figure 4 Omit-annealed $F_o - F_c$ map around AMPPCP at 5σ (a) and the hydrogen-bonding network around AMPPCP (b). AMPPCP is shown in ball-and-stick representation; the N- and P-domains are coloured light green and orange, respectively. Light-green broken lines in b show likely hydrogen bonds. A part of the N-domain is removed for clarity. Small spheres represent Mg^{2+} (green) and two water molecules (red), which coordinate to the Mg^{2+} , together with γ -phosphate, carboxyl groups of Asp 351 and Asp 703, and a carbonyl group of Thr 353 (orange broken lines in a and dark-green lines in b).

thereby forming a different interface with the A-domain.

The match of the P-domain with the catalytic domain in other members of the haloacid dehalogenase superfamily³⁵, such as phosphoserine phosphatase³⁶ and phosphoglucomutase³⁷, is rather poor in E1·2Ca²⁺ or E2(TG) form, but is much better in E1·AMPPCP (r.m.s.d. = 0.60 Å for 57 Cα atoms with phosphoserine phosphatase); the positions of Mg²⁺, coordinating residues and two water molecules are also nearly identical. These features are understandable, because Ca²⁺-ATPase requires much larger movements of the N-domain than the other members, such as phosphoserine phosphatase³⁶, while presumably using the same mechanism for phosphorylation and hydrolysis.

Positioning of the A-domain

The bending of the P-domain may also have a direct role in changing the orientation of the A-domain, which is tilted by ~30° around an

axis approximately parallel to the membrane (Figs 1 and 2). This can be regarded as a rigid-body movement around a hydrogen bond between main chains of two conserved residues, Gly 156 (A-domain) and Ala 725 (P-domain). This hydrogen bond is present in both E1·2Ca²⁺ and E1·AMPPCP (Fig. 5), but it is unlikely to be enough for linking the A- and P-domains together. In E1·2Ca²⁺, the A-domain appears to make a point contact with the P-domain (Fig. 5) and seems to be highly mobile: Gly 156–Lys 158, in a short loop sticking out from the A-domain, and Ala 725–Val 726 at the top of the P7 helix (Fig. 5b), are the only residues that come within a 4 Å distance.

This situation is not much different in E1·AMPPCP. Again, the interaction is limited to that around the P7 helix (Supplementary Fig. 2) but is more stable: the 'socket' on the P-domain is larger and one extra hydrogen bond appears to be formed (Supplementary Fig. 2). These apparently small differences in contacting residues, however, may be critical for changing the orientation of the A-domain, because the socket, the P7 helix, moves upwards and tilts so that the M2 side becomes higher up in the E1·2Ca²⁺ → E1·AMPPCP transition (Fig. 5b). Although small, this movement caused by the binding of the γ-phosphate and Mg²⁺ will be effective, because it occurs at the pivoting point.

The interface between the N- and A-domains is not complementary either (Fig. 6). There are only two contact spots located on either side of the apparent pivoting point (Fig. 1, dotted circles in red). Presumably the most important one is a kind of mechanical couple formed on the M3 side around a hydrogen bond between Thr 171 (A-domain) and Glu 486 (N-domain) (Fig. 6). On the M2 side, we can identify two salt bridges and three hydrogen bonds likely to be formed between the A- and N-domains. If the A-domain keeps to the same position as in E1·2Ca²⁺, when the N-domain inclines Thr 484 (N-domain) would collide with Thr 171 and push the M3 side of the A-domain. This might be another cause of the A-domain tilting; however, its contribution is difficult to evaluate, because the P-domain inclines and moves the A-domain at the same time.

Considering that the A-domain has to rotate horizontally by more than 90° during the E1P → E2P transition, it is understandable that the A–N interface is not complementary (Fig. 5). The structure around Thr 171 is interesting in this regard, because it may provide a solid pivot for such rotations, and because Thr 171

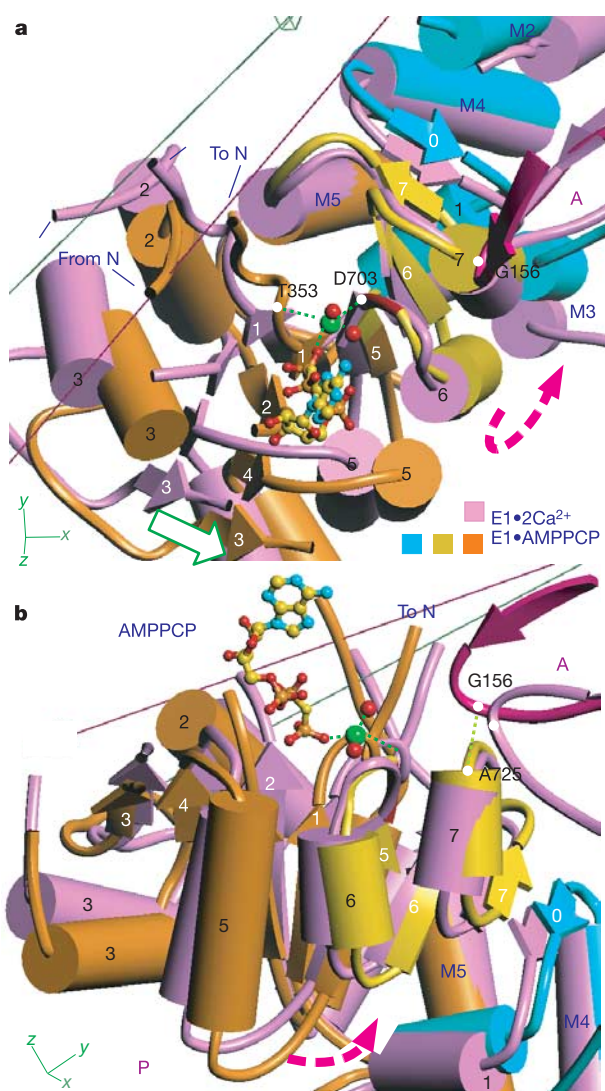


Figure 5 Superposition of the P-domain in E1·2Ca²⁺ and E1·AMPPCP, fitted with the 15 residues at the N-terminal end of the M5 helix. **a**, **b**, Top (**a**) and side (**b**) views. Colours representing E1·2Ca²⁺ and different regions of the P-domain in E1·AMPPCP, as recognized by Dyndom⁴⁵, are indicated. A part of the P-domain (containing helices 4 and 4') is removed. Small spheres represent Mg²⁺ (green) and two water molecules (red). Two long lines specify the positions of the axis of the N-domain inclination for E2(TG) (green) and E1·AMPPCP (purple) as determined by Dyndom⁴⁵. The green arrow in **a** indicates the movement of β-strands 1–4 in the E1·2Ca²⁺ → E1·AMPPCP transition, whereas those in red show the tilting of the α-helices 5–7.

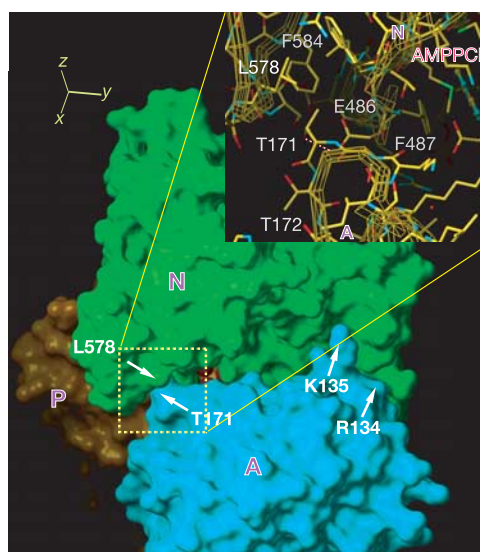


Figure 6 Water-accessible surface of the cytoplasmic domains in E1·AMPPCP showing the A–N interface. Inset is an enlarged view of the boxed area, where the A- and N-domains appear to be mechanically coupled.

interacts with Glu 486, the residue next to Phe 487, which is the primary residue for adenine binding. Hence, the binding and the release of ATP might directly affect the A–N interface. In fact the strand N β 7, on which Glu 486 and Phe 487 are located (Fig. 6 inset), shows the largest movement in the N-domain upon binding of AMPPCP.

Discussion

From the crystal structures it is clear that Ca²⁺-ATPase changes the orientation of the A-domain to regulate the cytoplasmic gate of the Ca²⁺-binding pathway by moving, primarily, the M1 helix. This means that the interfaces between the A-domain and the two other cytoplasmic domains are critically important and adjusted during the reaction cycle. Here, ATP works as the principal modifier. It brings the N-domain very close to the P-domain by directly bridging them, so that the N domain can ‘grip’ the M3 side of the A-domain. To achieve this, the γ -phosphate bends the P-domain to gain an extra $\sim 30^\circ$ of inclination. At the same time, ATP induces the binding of Mg²⁺, which bends the P-domain in a nearly orthogonal direction so that the M2 side is brought higher up. As a combined result of these interactions, the A-domain tilts by $\sim 30^\circ$ to pull up the M1–M2 helices and to strain the loop connecting to M3. This strain might be the driving force for another A-domain rotation that occurs in subsequent steps in the reaction cycle to open the luminal gate. The extensive hydrogen-bonding network formed around the ATP (analogue) might be required for sustaining this strained state. □

Methods

Crystallization

Ca²⁺-ATPase was prepared from rabbit hind-leg white muscle³⁸ and purified by affinity chromatography³⁹; for elution AMPPCP was used. Crystals of P₂₁ symmetry were prepared by dialysing purified enzyme (20 μ M) mixed with phosphatidylcholine in detergent octaethyleneglycol mono-*n*-decylether (C₁₂E₈, 2 mg ml^{−1}) against a buffer consisting of 20% glycerol, 12% PEG 400, 10 mM CaCl₂, 5 mM MgCl₂, 2.5 mM Na₃N₂, 2 μ g ml^{−1} butylhydroxytoluene, 0.2 mM dithiothreitol, 20 mM MES, pH 6.1, and 0.5 mM AMPPCP for about one month. Crystals were grown to 100 $\times$ 100 $\times$ 60 μ m and flash-frozen in cold nitrogen gas in a cold room. For making C2 crystals, 200 mM sodium propionate was included in the dialysis buffer.

Data collection

All the diffraction data were collected at BL41XU, SPring-8, using a MAR165 CCD detector at $\lambda = 0.72$ Å, and processed with Denzo and Scalepack⁴⁰. For refinement, diffraction data from two best crystals with P₂₁ symmetry were merged ($R_{\text{merge}} = 5.6\%$ with a redundancy of 6.7 and $I/\sigma = 27.0$) for 1/15.0 to 1/2.9 Å^{−1} (24.0%, 3.10 and 3.58, respectively, for the highest-resolution bin, 3.00–2.9 Å). Unit cell dimensions were $a = 90.9$, $b = 123.6$, $c = 151.8$ Å and $\beta = 107.2^\circ$. Two protein molecules were contained in the asymmetric unit.

Modelling and refinement

The crystal structure was determined by molecular replacement⁴⁴, first using the data from C2 crystals that contain only one protein molecule in an asymmetric unit. The starting model included the M4–M10 helices and the P-domain only. The N-domain was readily placed in the first map calculated from them, after solvent flattening with CNS⁴¹. The M3 helix and then the A-domain were added in the subsequent maps with no difficulty. Finally M1–M2 helices were constructed. The resulting atomic model was used as the starting model for refinement with the diffraction data from P₂₁ crystals. The diffraction data contained 70,272 reflections at 99.4% completeness (from 15 to 2.9 Å resolution). The atomic model consisting of 7,709 $\times$ 2 atoms (7,671 in the protein, 31 in AMPPCP, 3 ions and 4 water) was refined at a 2.9 Å resolution to R_{cryst} of 25.2% and R_{free} of 29.6% using CNS⁴¹ with a strong non-crystallographic symmetry restraint. r.m.s.d of the bond length and angle were 0.009 Å and 1.3°, respectively. The geometry of the model was examined with Procheck⁴²; 992 out of 994 residues were in the most favourable or favourable region in the Ramachandran plot. The secondary structure was assigned with DSSP⁴³. (More details are given in Supplementary Methods.)

Received 23 March; accepted 19 May 2004; doi:10.1038/nature02680.

Published online 30 June 2004.

- Möller, J. V., Juul, B. & le Maire, M. Structural organization, ion transport, and energy transduction of P-type ATPases. *Biochim. Biophys. Acta* **1286**, 1–51 (1996).
- Post, R. L., Hegyváry, C. & Kume, S. Activation by adenosine triphosphate in the phosphorylation kinetics of sodium and potassium ion transport adenosine triphosphatase. *J. Biol. Chem.* **247**, 6530–6540 (1972).
- Albers, R. W. Biochemical aspects of active transport. *Annu. Rev. Biochem.* **36**, 727–756 (1967).
- de Meis, L. & Vianna, A. L. Energy interconversion by the Ca²⁺-dependent ATPase of the sarcoplasmic

reticulum. *Annu. Rev. Biochem.* **48**, 275–292 (1979).

- Mintz, E. & Guillain, F. Ca²⁺ transport by the sarcoplasmic reticulum ATPase. *Biochim. Biophys. Acta* **1318**, 52–70 (1997).
- Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405**, 647–655 (2000).
- Toyoshima, C. & Nomura, H. Structural changes in the calcium pump accompanying the dissociation of calcium. *Nature* **418**, 605–611 (2002).
- Xu, C., Rice, W. J., He, W. & Stokes, D. L. A structural model for the catalytic cycle of Ca²⁺-ATPase. *J. Mol. Biol.* **316**, 201–211 (2002).
- MacLennan, D. H., Rice, W. J. & Green, N. M. The mechanism of Ca²⁺ transport by sarco(endo)plasmic reticulum Ca²⁺-ATPases. *J. Biol. Chem.* **272**, 28815–28818 (1997).
- MacLennan, D. H., Brandl, C. J., Korczak, B. & Green, N. M. Amino-acid sequence of a Ca²⁺ + Mg²⁺-dependent ATPase from rabbit muscle sarcoplasmic reticulum, deduced from its complementary DNA sequence. *Nature* **316**, 696–700 (1985).
- Brandl, C. J., deLeon, S., Martin, D. R. & MacLennan, D. H. Adult forms of the Ca²⁺ ATPase of sarcoplasmic reticulum. Expression in developing skeletal muscle. *J. Biol. Chem.* **262**, 3768–3774 (1987).
- Sagara, Y. & Inesi, G. Inhibition of the sarcoplasmic reticulum Ca²⁺ transport ATPase by thapsigargin at subnanomolar concentrations. *J. Biol. Chem.* **266**, 13503–13506 (1991).
- Toyoshima, C., Nomura, H. & Sugita, Y. Structural basis of ion pumping by Ca²⁺-ATPase of sarcoplasmic reticulum. *FEBS Lett.* **555**, 106–110 (2003).
- Brünger, A. T. Extension of molecular replacement: a new search strategy based on Patterson correlation refinement. *Acta Crystallogr. A* **46**, 46–57 (1990).
- Danko, S., Yamasaki, K., Daiho, T., Suzuki, H. & Toyoshima, C. Organization of cytoplasmic domains of sarcoplasmic reticulum Ca²⁺-ATPase in E₁P and E₁ATP states: a limited proteolysis study. *FEBS Lett.* **505**, 129–135 (2001).
- Juul, B. *et al.* Do transmembrane segments in proteolyzed sarcoplasmic reticulum Ca²⁺-ATPase retain their functional Ca²⁺ binding properties after removal of cytoplasmic fragments by proteinase K? *J. Biol. Chem.* **270**, 20123–20134 (1995).
- Ma, H., Inesi, G. & Toyoshima, C. Substrate-induced conformational fit and headpiece closure in the Ca²⁺ ATPase (SERCA). *J. Biol. Chem.* **278**, 28938–28943 (2003).
- Andersen, J. P. & Vilsen, B. Amino acids Asn796 and Thr799 of the Ca²⁺-ATPase of sarcoplasmic reticulum bind Ca²⁺ at different sites. *J. Biol. Chem.* **269**, 15931–15936 (1994).
- Zhang, Z. *et al.* Detailed characterization of the cooperative mechanism of Ca²⁺ binding and catalytic activation in the Ca²⁺ transport (SERCA) ATPase. *Biochemistry* **39**, 8758–8767 (2000).
- Orlowski, S. & Champeil, P. Kinetics of calcium dissociation from its high-affinity transport sites on sarcoplasmic reticulum ATPase. *Biochemistry* **30**, 352–361 (1991).
- Vilsen, B. & Andersen, J. P. Mutation to the glutamate in the fourth membrane segment of Na⁺, K⁺-ATPase and Ca²⁺-ATPase affects cation binding from both sides of the membrane and destabilizes the occluded enzyme forms. *Biochemistry* **37**, 10961–10971 (1998).
- Inesi, G., Ma, H., Lewis, D. & Xu, C. Ca²⁺ occlusion and gating function of Glu309 in the fluoroaluminate analog of the Ca²⁺-ATPase phosphoenzyme intermediate. *J. Biol. Chem.* (in the press).
- Einholm, A. P., Vilsen, B. & Andersen, J. P. Importance of transmembrane segment M1 of the sarcoplasmic reticulum Ca²⁺-ATPase in Ca²⁺ occlusion and phosphoenzyme processing. *J. Biol. Chem.* **279**, 15888–15896 (2004).
- Bilwes, A. M., Quezada, C. M., Croal, L. R., Crane, B. R. & Simon, M. I. Nucleotide binding by the histidine kinase CheA. *Nature Struct. Biol.* **8**, 353–360 (2001).
- Hilge, M. *et al.* ATP-induced conformational changes of the nucleotide-binding domain of Na,K-ATPase. *Nature Struct. Biol.* **10**, 468–474 (2003).
- McIntosh, D. B., Woolley, D. G., Vilsen, B. & Andersen, J. P. Mutagenesis of segment ⁴⁸⁷Phe-Ser-Arg-Asp-Arg-Lys⁴⁹² of sarcoplasmic reticulum Ca²⁺-ATPase produces pumps defective in ATP binding. *J. Biol. Chem.* **271**, 25778–25789 (1996).
- Cappello, V., Tramontano, A. & Koch, U. Classification of proteins based on the properties of the ligand-binding site: the case of adenine-binding proteins. *Proteins* **47**, 106–115 (2002).
- Liu, M. & Barth, A. Mapping interactions between the Ca²⁺-ATPase and its substrate ATP with infrared spectroscopy. *J. Biol. Chem.* **278**, 10112–10118 (2003).
- Clarke, D. M., Loo, T. W. & MacLennan, D. H. Functional consequences of alterations to amino acids located in the nucleotide binding domain of the Ca²⁺-ATPase of sarcoplasmic reticulum. *J. Biol. Chem.* **265**, 22223–22227 (1990).
- Maruyama, K. *et al.* Functional consequences of alterations to amino acids located in the catalytic center (isoleucine 348 to threonine 357) and nucleotide-binding domain of the Ca²⁺-ATPase of sarcoplasmic reticulum. *J. Biol. Chem.* **264**, 13038–13042 (1989).
- McIntosh, D. B., Woolley, D. G., MacLennan, D. H., Vilsen, B. & Andersen, J. P. Interaction of nucleotides with Asp³⁵¹ and the conserved phosphorylation loop of sarcoplasmic reticulum Ca²⁺-ATPase. *J. Biol. Chem.* **274**, 25227–25236 (1999).
- Clausen, J. D., McIntosh, D. B., Vilsen, B., Woolley, D. G. & Andersen, J. P. Importance of conserved N-domain residues Thr441, Glu442, Lys515, Arg560, and Leu562 of sarcoplasmic reticulum Ca²⁺-ATPase for MgATP binding and subsequent catalytic steps. Plasticity of the nucleotide-binding site. *J. Biol. Chem.* **278**, 20245–20258 (2003).
- Jacobsen, M. D., Pedersen, P. A. & Jørgensen, P. L. Importance of Na,K-ATPase residue α 1-Arg544 in the segment Arg544-Asp567 for high-affinity binding of ATP, ADP, or MgATP. *Biochemistry* **41**, 1451–1456 (2002).
- McIntosh, D. B. The ATP binding sites of P-type ion transport ATPases. *Adv. Mol. Cell Biol.* **23A**, 33–99 (1998).
- Aravind, L., Galperin, M. Y. & Koonin, E. V. The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold. *Trends Biochem. Sci.* **23**, 127–129 (1998).
- Wang, W. *et al.* Structural characterization of the reaction pathway in phosphoserine phosphatase: crystallographic ‘snapshots’ of intermediate states. *J. Mol. Biol.* **319**, 421–431 (2002).
- Lahiri, S. D., Zhang, G., Dunaway-Mariano, D. & Allen, K. N. Caught in the act: the structure of phosphorylated β -phosphoglucomutase from *Lactococcus lactis*. *Biochemistry* **41**, 8351–8359 (2002).
- Eletr, S. & Inesi, G. Phospholipid orientation in sarcoplasmic membranes: spin-label ESR and proton NMR studies. *Biochim. Biophys. Acta* **282**, 174–179 (1972).
- Coll, R. J. & Murphy, A. J. Purification of the CaATPase of sarcoplasmic reticulum by affinity chromatography. *J. Biol. Chem.* **259**, 14249–14254 (1984).

40. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–325 (1997).
41. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
42. Collaborative Computational Project No. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
43. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
44. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
45. Hayward, S. Structural principles governing domain motions in proteins. *Proteins* **36**, 425–435 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Tsuda for help in many aspects of this work; M. Kawamoto and H. Sakai for data collection at BL41XU of SPring-8; N. Miyashita for making many movies; and Y. Ohuchi for computer programs. We are grateful to D.B. McIntosh for help in improving the manuscript and G. Inesi for communicating unpublished results to us. This work was supported in part by a Creative Science Project Grant from the Ministry of Education, Culture, Sports, Science and Technology, the Japan New Energy and Industry Technology Development Organization, and the Human Frontier Science Program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.T. (ct@iam.u-tokyo.ac.jp). The atomic coordinates are deposited in the PDB under accession code 1VFP.

Solar chromospheric spicules from the leakage of photospheric oscillations and flows

Bart De Pontieu¹, Robert Erdélyi² & Stewart P. James²

¹Lockheed Martin Solar & Astrophysics Laboratory, 3251 Hanover Street, Org. ADBS, Building 252, Palo Alto, California 94304, USA

²Solar Physics and Upper-Atmosphere Research Centre, Department of Applied Mathematics, University of Sheffield, Hicks Building, Hounsfield Road, Sheffield S3 7RH, UK

Spicules are dynamic jets propelled upwards (at speeds of $\sim 20 \text{ km s}^{-1}$) from the solar 'surface' (photosphere) into the magnetized low atmosphere of the Sun^{1–3}. They carry a mass flux of 100 times that of the solar wind into the low solar corona⁴. With diameters close to observational limits ($< 500 \text{ km}$), spicules have been largely unexplained⁵ since their discovery in 1877⁶: none of the existing models³ can account simultaneously for their ubiquity, evolution, energetics and recently discovered periodicity⁶. Here we report a synthesis of modelling and high-spatial-resolution observations in which numerical simulations driven by observed photospheric velocities directly reproduce the observed occurrence and properties of individual spicules. Photospheric velocities are dominated by convective granulation (which has been considered before for spicule formation^{7–11}) and by p-modes (which are solar global resonant acoustic oscillations visible in the photosphere as quasi-sinusoidal velocity and intensity pulsations). We show that the previously ignored p-modes are crucial: on inclined magnetic flux tubes, the p-modes leak sufficient energy from the global resonant cavity into the chromosphere to power shocks that drive upward flows and form spicules.

Spicules are short-lived (~ 5 – 10 min) elongated features seen at the solar limb. On the disk they are historically called mottles in quiet Sun regions and fibrils in active regions¹². Spicules are ubiquitous, numbering $\sim 100,000$ at any time¹, and usually occur in association with intense photospheric magnetic flux tubes¹³ (Fig. 1). Directly above active region plage, they appear as short fibrils that typically reach $3,000$ – $4,000 \text{ km}$ height above the photosphere (Fig. 1), compared to $\sim 5,000 \text{ km}$ or more in quiet Sun^{1,2}. Recently, we found that the occurrence of these fibrils often shows periodicity with a peak around 5 min (ref. 6). Our wavelet analysis of high-resolution images obtained with the new 1 m Swedish Solar Telescope¹⁴ reveals how pervasive these periodic signals are, and how they are related to individual spicules (Fig. 1). Intriguing similarities in period and wave train duration suggest that periodic spicules are related to p-modes. Such a connection has not been considered by existing theories, because observational evidence was lacking and p-modes are evanescent in the upper photosphere.

It is the leakage of p-mode oscillations into the chromosphere, and the nonlinear evolution of photospheric flows and tunnelled oscillations into the inhomogeneous solar atmosphere, that hold the key to the spicule phenomenon. We consider a numerical model¹⁵ of an initially hydrostatic atmosphere (VAL III¹⁶) confined within a rigid flux tube, which expands with height¹³ as the magnetic field strength changes from $1,600 \text{ G}$ in the photosphere to 120 G in the low corona—appropriate values for active region plage¹⁷. The base of the flux tube is driven by photospheric velocities observed with SOHO/MDI¹⁸, for a range of inclination angles θ between the flux tube axis and the vertical. The time series of MDI velocities is determined by granular convective flows and a dominant oscillatory signal caused by the superposition of p-modes¹⁹.

The influence of granular flows on spicule formation has been studied extensively in the so-called rebound shock model^{8–11}. Although some features of this model appear here, the rebound shock model cannot account for the periodicity, evolution and energetics of the spicules in Fig. 2: it considers only short-lived ($\sim 90 \text{ s}$) velocity pulses, and ignores the dominant oscillatory component in the photospheric velocity signal. Previous studies of the effects of these photospheric oscillations on the chromosphere have largely ignored their potential for driving mass flows, focusing instead on chromospheric heating and neglecting the role of flux tubes and their expansion and non-verticality^{20,21}, or the presence of a hot chromosphere²²: all of these are crucial in spicule formation.

Our simulations show that the photospheric flows and oscillations leak into the chromosphere, where they develop nonlinearly into upward propagating acoustic shocks. These shocks impart upward momentum to the chromosphere (~ 20 – 40 km s^{-1}) and lead to significant periodic excursions of the top of the chromosphere, causing spicule formation (Fig. 2). With observed photospheric velocities driving our simulations, we calculate the height of spicules occurring above two arbitrary though widely separated locations (a , b) in different patches of plage. We find, to our

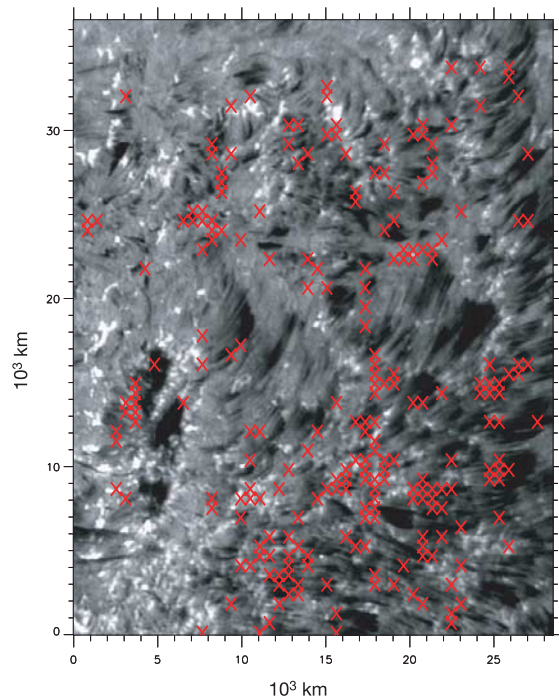


Figure 1 High-resolution image showing the ubiquity and periodicity of spicules. Spicules form cool and dense intrusions ($\sim 8,000 \text{ K}$ and $\sim 5 \times 10^{16} \text{ m}^{-3}$)^{1,2,28} into the hot and rarefied corona ($\sim 1 \text{ MK}$ and $\sim 10^{15} \text{ m}^{-3}$). Spicules are visible as dark features with spatial extent of $2,000$ – $3,000 \text{ km}$ in this image (obtained from the Swedish 1-m Solar Telescope¹⁴) of active region 10380 taken on 16 June 2003 in the blue wing (at $\sim 700 \text{ mÅ}$) of the chromospheric $\text{H}\alpha$ line using adaptive optics. Tickmarks have $1,000 \text{ km}$ spacing. Most spicules occur close to magnetic flux tubes, visible here as a network of bright points that together form a magnetic plage region. Close by are some small sunspots (at $x = 0$ – 7 , $y = 8$ – 18). Red crosses show locations where spicular periodicity around 5 min was observed in the Doppler-shifted wings of the chromospheric $\text{H}\alpha$ line during a time span of 1 h . Analysis of time series of these high-resolution data reveals that the observed periodicity is caused by the repeated appearance of spicules at intervals of $350 \pm 60 \text{ s}$.

knowledge for the first time, that the predicted temporal evolution of spicule height in both cases matches the observed occurrence of different spicules (Fig. 2) remarkably well for a time range of 2,000 s (location *a*) and 2,500 s (location *b*), for flux tube inclinations $\theta = 0^\circ$ (*a*) and $\theta = 50^\circ$ (*b*). The different inclination angles are compatible with the location of *a* in a strong plage region and of *b* at the outer edge of plage where highly inclined field lines are prominent. The time difference between the observed spicule occurrence and the predicted maxima of spicule height, for

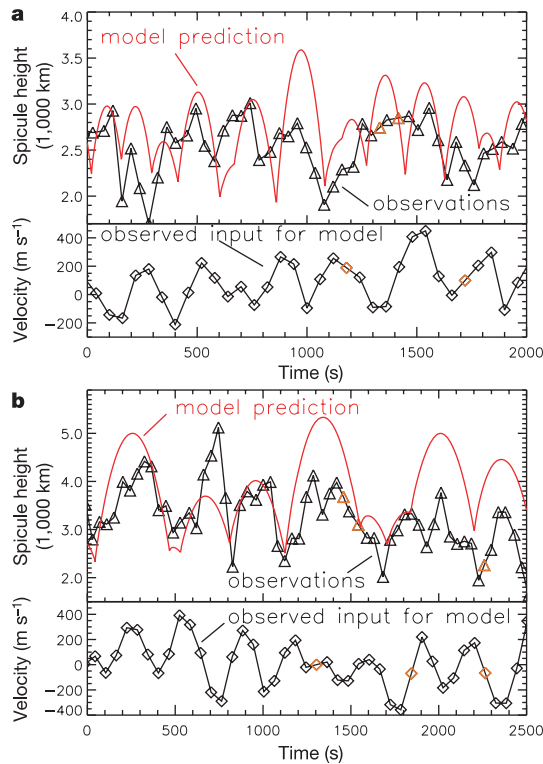


Figure 2 Comparison between spicule occurrence observed on the Sun and spicule height obtained from numerical simulations that are driven by observed velocities in the photosphere. Panel **a** shows when spicules occurred on 4 June 1999 (full black line, triangles) at one location (*a*) on the Sun, as deduced from temporal variations of the extreme ultraviolet (EUV) brightness I_{171} (TRACE²⁶ 171 Å) of the upper transition region in a 725 km × 725 km region above active region plage. The full black line is proportional to $-c I_{171}$, with c an arbitrary scaling factor. The time of EUV brightness variations is highly correlated with spicule occurrence^{6,29}, when cold spicules protrude over neighbouring low-lying hot upper transition region plasma, they absorb EUV photons emitted by the hot plasma, causing a dip in EUV emission (and an increase in $-I_{171}$). The observed spicule occurrence (times of maxima of $-I_{171}$) fits remarkably well with the temporal evolution of the spicule height (full red line) predicted by a one-dimensional nonlinear numerical simulation of a vertical flux tube ($\theta = 0^\circ$) that is driven by an artificial force based on photospheric velocity data from SOHO/MDI¹⁸ (lower panel) observations at the same location *a*. The absolute amplitude of the EUV brightness variations is not a good proxy for spicule height: spicules have a wide range of diameters so that the amount of EUV absorption can vary considerably from event to event, leading to a variable correlation between the amplitude of variations in $-I_{171}$ and modelled spicule heights. Orange triangles and diamonds show where, respectively, TRACE or MDI data were missing, and interpolation was used. Modelled spicule heights are less reliable within 250–350 s of missing MDI data. All transitory effects in the simulation have decayed by $t = 0$ s. **b**, same as in **a**, but for a flux tube with inclination angle $\theta = 50^\circ$, at a different location (*b*) in a nearby plage region.

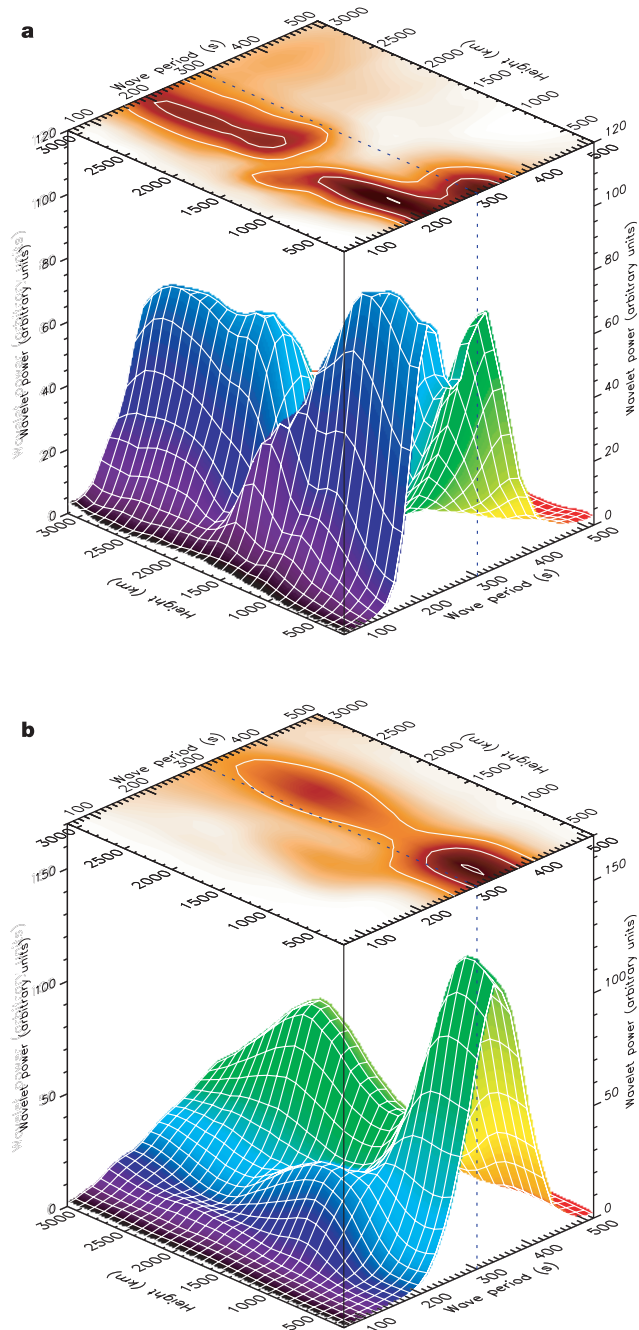


Figure 3 Leakage of 'evanescent' p-mode power from photosphere into chromosphere. Distribution of wavelet power (in arbitrary units, independent for each height) during the time periods shown in Figs 2 and 4 (for cases **a** and **b**, respectively $\theta = 0^\circ$ and 50°) as a function of wave period (x) from 100 to 500 s for different heights in the solar atmosphere (y) from 200 to 3,000 km above the photosphere. As a reference, dashed blue lines show the typical peak of photospheric p-mode power at 300 s. Inclined flux tubes (**b**) show an increased acoustic cut-off period P_c , allowing enhanced leakage and direct propagation of normally evanescent ($P > 300$ s) p-modes through the temperature minimum into the chromosphere. Vertical flux tubes (**a**) have $P_c \approx 200$ s in the upper photosphere, and allow some leakage of p-modes with 300 s periods. However, oscillations with lower periods (< 250 s), if present in the photospheric velocity signal, can leak more or propagate and grow with height to dominate chromospheric dynamics. Generally, spicule periodicity depends on the frequencies present in the photospheric velocity signal, and on the inclination angle of the flux tube. Even relatively minor 200 s power in the photosphere can grow with height to dominate the chromosphere if the much stronger photospheric 300 s power cannot tunnel significantly. However, flux tubes with only modest inclinations from the vertical allow significant leakage of p-modes so that their typical period of 300 s can dominate the chromospheric power spectrum, and spicule formation (see also Fig. 4).

locations *a* and *b* respectively, is on average 20 ± 12 s and 26 ± 30 s, a fraction of the spicule lifetime (~ 300 s), and well within the ± 35 s and ± 60 s time range during which the maximum height remains approximately constant.

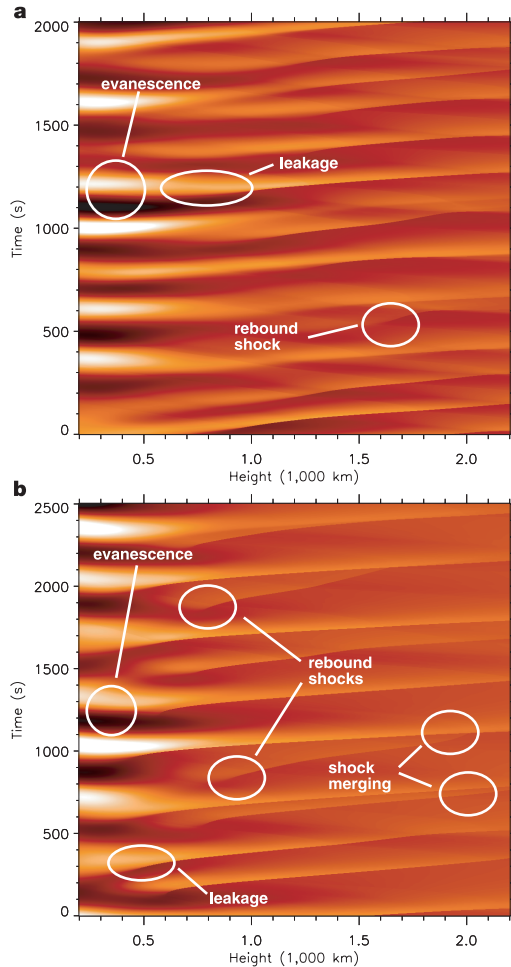


Figure 4 Mass-weighted vertical velocity along the flux tube as a function of height and time. The velocity (upwards is white) is calculated for numerical simulations **a**, $\theta = 0^\circ$ and **b**, $\theta = 50^\circ$ (Figs 2, 3). Photospheric oscillations, evanescent at low heights, tunnel into the chromosphere where they can propagate (for $x > 0.7$ in **a**, $x > 0.5$ in **b**), steepen into shocks, and drive upward flows that form spicules. These shocks leave behind an oscillating wake in the chromosphere ($x > 0.7$), which sometimes nonlinearly develops into a rebound shock⁸. When rebound shocks merge with other shocks, they are strengthened and form taller spicules. Shock merging occurs more often when leaked p-modes 'resonate' with the natural response of the middle chromosphere, that is, when the dominant period of the p-modes is close to that of the oscillating wake. The period of the oscillating wake is given by the acoustic cut-off period of the middle chromosphere^{8,30}, which at 250–350 s (depending on flux tube inclination) is remarkably close to the dominant period of p-modes. The strength of the shocks is dependent on the radiative losses of our model: a newtonian cooling term $\sim (T - T_0)/t$, with T the temperature and T_0 the temperature of the undisturbed atmosphere, with time constant $t \approx 250$ s. Such a cooling term is only an approximation to the real radiative losses in the optically thick chromosphere. However, it avoids the exaggerated spicule temperatures and heights associated with an adiabatic approach^{8,9}, as well as the very large radiative losses of the optically thin approximation¹¹, and leads to spicule temperatures of 8,000–12,000 K (compatible with observations^{1,2,28}) and heights of $\sim 4,000$ km (compatible with spicule heights in Fig. 1). Whatever the real radiative losses are, their influence on the strength of the shocks leads to somewhat different spicule heights, and not as much to changes in timing of the shocks and spicules. The timing of the spicules is determined mostly by the photospheric driver.

How can photospheric oscillations drive upper chromospheric jets? In the upper photosphere, p-modes are evanescent and cannot propagate upwards through the temperature minimum¹⁹: under classical assumptions ($\gamma = 5/3$, $\theta = 0^\circ$), their period P (≈ 200 –450 s) is above the local acoustic cut-off period, $P_c \approx 180$ s. Here $P_c = 2\pi/\Omega \sim \sqrt{T}/(\sqrt{\gamma}\cos\theta)$, where the acoustic cut-off frequency $\Omega = \gamma g/(2c)$, with the ratio of specific heats γ , effective gravity $g = g_0 \cos\theta$ ($g_0 = 274 \text{ m s}^{-2}$), inclination angle θ and sound speed $c = \sqrt{\gamma RT/\mu}$ with gas constant R , temperature T and mean molecular weight μ . However, as we observe in this model (Figs 3, 4), evanescence does not preclude significant tunnelling of non-propagating wave energy through the temperature minimum into the hot chromosphere of a flux tube, where propagation is once again possible (Fig. 3) because of higher temperatures ($P_c > 300$ s). More importantly, the non-verticality of flux tubes ($P_c \sim 1/\cos\theta$) dramatically increases tunnelling, and can even lead to propagation of p-modes along the flux tube's inclined field lines. The dependence on inclination has not been previously recognized in this context, and is responsible for the dominant periodicity of spicule formation in our simulations (Figs 3, 4).

The leaked photospheric velocity signals steepen into shocks as the atmospheric density rapidly decreases with height. The passage of these shocks leaves an oscillating wake in the chromosphere, which sometimes nonlinearly develops into a rebound shock⁸ (Fig. 4). However, unlike in the rebound shock model^{8–11}, these shocks are not a necessary condition for spicule formation, with 12 out of 19 spicules in our simulations showing no evidence of such shocks. It is the interaction of our continuous photospheric driver (consisting of both granular and oscillatory signals) with the chromospheric response that is crucial for the good match with observations.

Our model yields a straightforward explanation for the ubiquity of spicules. Flux tubes provide a necessary condition for spicule formation: first, their intense magnetic field guides the photospheric signal limiting the three-dimensional wave scattering that occurs in a non-magnetic environment; and second, the high temperature of the flux tube chromosphere, which may not be present in a non-magnetic environment²³, guarantees chromospheric propagation of tunnelled p-modes. The photospheric diameter of flux tubes (< 100 km) is a natural lower limit to spicular diameters and explains why spicules are much narrower than the horizontal coherence length of p-modes ($\sim 8,000$ km). Spicule formation nonlinearly depends on highly localized properties: (1) the local strength of the photospheric velocity signal, which is determined by p-modes, as well as flows on scales < 500 km caused by granules, (2) the temperature and partial ionization from photosphere through chromosphere, and (3) the flux tube inclination.

Generally, the temperature stratification of the atmosphere acts as a high-pass filter, with the inclination of the magnetic field decreasing the low-frequency cut-off of the filter, and determining which, if any, constituents of the photospheric signal will leak or propagate, and drive chromospheric flows. Higher inclinations cause enhanced leakage of p-modes into the chromosphere through a reduced layer of evanescence, and thus a stronger likelihood of 'resonance' between the leaked p-modes and the natural response of the middle chromosphere. This can explain why short fibrils (Fig. 1) and quiet Sun mottles^{1,2} typically show significant deviations from the vertical, and why much of the 5 min power in spicules has been observed towards the periphery of plage regions⁶, where field lines are more inclined. As a flux tube expands with height, individual field lines connect to different locations, so a wide range of inclinations may be presented to photospheric velocities. Only on some of these field lines will the filtered signal be such that periodic or any kind of spicules develop, depending on the power spectrum of the photospheric signal. This mechanism

may also explain the larger extent of quiet Sun spicules^{1,2} where p-mode power and granular flows are stronger by up to 50% and magnetic fields are more inclined owing to the presence of opposite polarity^{24,25}.

A natural consequence of our model is that the quasi-periodic shocks driving the spicules propagate upwards into the low corona, where they may lead to intensity oscillations with properties that are similar to those of longitudinal oscillations observed by TRACE²⁶ in 1 MK coronal loops²⁷. □

Received 31 March; accepted 14 June 2004; doi:10.1038/nature02749.

1. Beckers, J. M. Solar spicules. *Sol. Phys.* **3**, 367–433 (1968).
2. Beckers, J. M. Solar spicules. *Annu. Rev. Astron. Astrophys.* **10**, 73–100 (1972).
3. Sterling, A. C. Solar spicules: a review of recent models and targets for future observations. *Sol. Phys.* **196**, 79–111 (2000).
4. Withbroe, G. L. The role of spicules in heating the solar atmosphere: implications of EUV observations. *Astrophys. J.* **267**, 825–836 (1983).
5. Foukal, P. *Solar Astrophysics* 10–11 (John Wiley & Sons, New York, 1990).
6. De Pontieu, B., Erdélyi, R. & de Wijn, A. G. Intensity oscillations in the upper transition region above active region plage. *Astrophys. J.* **595**, L63–L66 (2003).
7. Roberts, B. The resonant response to granular buffeting. *Sol. Phys.* **61**, 23–34 (1979).
8. Hollweg, J. V. On the origin of solar spicules. *Astrophys. J.* **257**, 345–353 (1982).
9. Sterling, A. C. & Hollweg, J. V. The rebound shock model for solar spicules—Dynamics at long times. *Astrophys. J.* **327**, 950–963 (1988).
10. Sterling, A. C. & Hollweg, J. V. A rebound shock model for solar fibrils. *Astrophys. J.* **343**, 985–993 (1989).
11. Sterling, A. C. & Mariska, J. T. Numerical simulations of the rebound shock model for solar spicules. *Astrophys. J.* **349**, 647–655 (1990).
12. Suematsu, Y., Wang, H. & Zirin, H. High-resolution observation of disk spicules: I. Evolution and kinematics of spicules in the enhanced network. *Astrophys. J.* **450**, 411–421 (1995).
13. Spruit, H. C. & Roberts, B. Magnetic flux tubes on the sun. *Nature* **304**, 401–406 (1983).
14. Scharmer, G. B., Bjelksjö, K., Korhonen, T. K., Lindberg, B. & Pettersson, B. in *Innovative Telescopes and Instrumentation for Solar Astrophysics* (eds Keil, S. & Avakyan, S.) 341–350 (Proc. SPIE, Vol. 4853, SPIE—The International Society for Optical Engineering, Hawaii, 2003).
15. James, S. P., Erdélyi, R. & De Pontieu, B. Can ion-neutral damping help to form spicules? *Astron. Astrophys.* **406**, 715–724 (2003).
16. Vernazza, J. E., Avrett, E. H. & Loeser, R. Structure of the solar chromosphere. III—Models of the EUV brightness components of the quiet-sun. *Astrophys. J. Suppl.* **45**, 635–725 (1981).
17. Schrijver, C. J. & Harvey, K. L. The photospheric magnetic flux budget. *Sol. Phys.* **150**, 1–18 (1994).
18. Scherrer, P. H. *et al.* The Solar Oscillations Investigation—Michelson Doppler Imager. *Sol. Phys.* **162**, 129–188 (1995).
19. Christensen-Dalsgaard, J., Gough, D. & Toomre, J. Seismology of the Sun. *Science* **229**, 923–931 (1985).
20. Cheng, Q.-Q. & Yi, Z. Oscillations in the solar atmosphere: a result of hydrodynamical simulations. *Astron. Astrophys.* **313**, 971–978 (1996).
21. Sutmann, G. & Ulmschneider, P. Acoustic wave propagation in the solar atmosphere: II. Nonlinear response to adiabatic wave excitation. *Astron. Astrophys.* **294**, 241–251 (1995).
22. Museliak, Z. E. & Ulmschneider, P. Atmospheric oscillations in solar magnetic flux tubes. I. Excitation by longitudinal waves and random pulses. *Astron. Astrophys.* **400**, 1057–1064 (2003).
23. Carlsson, M. & Stein, R. F. Does a nonmagnetic solar chromosphere exist? *Astrophys. J.* **440**, L29–L32 (1995).
24. Schrijver, C. J. & Title, A. M. The magnetic connection between the solar photosphere and the corona. *Astrophys. J.* **597**, L165–L168 (2003).
25. Schrijver, C. J. *et al.* A new view of the solar outer atmosphere by the transition region and coronal explorer. *Sol. Phys.* **187**, 261–302 (1999).
26. Handy, B. N. *et al.* The transition region and coronal explorer. *Sol. Phys.* **187**, 229–260 (1999).
27. De Moortel, I., Ireland, J., Walsh, R. W. & Hood, A. W. Longitudinal intensity oscillations in coronal loops observed with TRACE — I. Overview of measured parameters. *Sol. Phys.* **209**, 61–88 (2002).
28. Lindsey, C. *et al.* Extreme-infrared brightness profile of the solar chromosphere obtained during the total eclipse of 1991. *Nature* **358**, 308–310 (1992).
29. De Pontieu, B., Tarbell, T. D. & Erdélyi, R. Correlations on arcsecond scales between chromospheric and transition region emission in active regions. *Astrophys. J.* **590**, 502–518 (2003).
30. Rae, I. C. & Roberts, B. Pulse propagation in a magnetic flux tube. *Astrophys. J.* **256**, 761–767 (1982).

Acknowledgements This work was supported by NASA, the UK Particle Physics Research Council (PPARC) and NSF Hungary. The Swedish Solar Telescope is operated on the island of La Palma by the Royal Swedish Academy of Sciences in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias. R.E. thanks M. Kéray for encouragement. We thank C.J. Schrijver, T. Tarbell, M. DeRosa and A. Title for discussions, and M. Carlsson for pointing out the importance of 3 min oscillations.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.D.P. (bdp@lmsal.com).

A universal scaling relation in high-temperature superconductors

C. C. Homes¹, S. V. Dordevic¹, M. Strongin¹, D. A. Bonn², Ruixing Liang², W. N. Hardy², Seiki Komiya³, Yoichi Ando³, G. Yu⁴, N. Kaneko^{5*}, X. Zhao⁵, M. Greven^{5,6}, D. N. Basov⁷ & T. Timusk⁸

¹Department of Physics, Brookhaven National Laboratory, Upton, New York 11973, USA

²Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia V6T 2A6, Canada

³Central Research Institute of Electric Power Industry, Komae, Tokyo 201-8511, Japan

⁴Department of Physics, Stanford University, Stanford, California 94305, USA

⁵Stanford Synchrotron Radiation Laboratory, Stanford, California 94309, USA

⁶Department of Applied Physics, Stanford University, Stanford, California 94305, USA

⁷Department of Physics, University of California at San Diego, La Jolla, California 92093, USA

⁸Department of Physics and Astronomy, McMaster University, Hamilton, Ontario L8S 4M1, Canada

* Present address: National Institute of Advanced Industrial Science and Technology, Tsukuba Central 2-2, Tsukuba, Ibaraki 305-8568, Japan

Since the discovery of superconductivity at elevated temperatures in the copper oxide materials¹ there has been a considerable effort to find universal trends and correlations amongst physical quantities, as a clue to the origin of the superconductivity. One of the earliest patterns that emerged was the linear scaling of the superfluid density (ρ_s) with the superconducting transition temperature (T_c), which marks the onset of phase coherence. This is referred to as the Uemura relation², and it works reasonably well for the underdoped materials. It does not, however, describe optimally doped (where T_c is a maximum) or overdoped materials³. Similarly, an attempt to scale the superfluid density with the d.c. conductivity (σ_{dc}) was only partially successful⁴. Here we report a simple scaling relation ($\rho_s \propto \sigma_{dc} T_c$, with σ_{dc} measured at approximately T_c) that holds for all tested high- T_c materials. It holds regardless of doping level, nature of dopant (electrons versus holes), crystal structure and type of disorder⁵, and direction (parallel or perpendicular to the copper-oxygen planes).

We first demonstrate scaling for the a - b plane (that is, parallel to the copper-oxygen planes) properties^{6–12} of single- and double-layer copper oxide materials (Supplementary Table 1), as well as for the conventional metals^{13,14} Nb and Pb (elemental superconductors with relatively high values of T_c). The values for ρ_s and σ_{dc} are obtained simultaneously from studies of the reflectance of these materials. The results for the scaling relation are shown on a log-log plot in Fig. 1. The dashed line is a linear fit to the data, while the dotted lines form what are effectively an upper and lower bound for the data; this is described by $\rho_s = (120 \pm 25)\sigma_{dc} T_c$ (where ρ_s is in cm^{-2} , σ_{dc} is in $\Omega^{-1} \text{cm}^{-1}$, and T_c is in K). The remarkable result contained in this plot is that within error all of these points fall onto a single line with a slope of unity. This is significant, as the optimally and overdoped materials, which fall well off of the Uemura plot, now scale with the underdoped materials onto a single line.

We also searched for scaling relations along the poorly conducting c axis, where the charge transport is thought to be incoherent¹⁵. Previous work focused on scaling between ρ_s and σ_{dc} only^{16,17}. Whereas this approach yields good results for the underdoped materials, in a fashion reminiscent of the Uemura plot, significant deviations from linearity are encountered for optimally and overdoped materials; this was thought to signal the onset of more conventional three-dimensional behaviour. Figure 2 demonstrates

that the c -axis data^{11,17–20} for all of the single and double-layer materials (Supplementary Table 2) are again well described by a line with slope of unity. What is perhaps most remarkable is that the a - b -plane and c -axis results may all be described by the same universal line shown in Fig. 2, even though the two results correspond to very different ranges of ρ_s . The combined data span nearly five orders of magnitude, from the insulating behaviour along the c axis in the underdoped systems, to the metallic behaviour in the a - b planes of the overdoped copper oxides.

The scaling relation for the a - b planes can be interpreted in a number of different ways. One of the most direct is the assumption that all of the spectral weight (the area obtained from the integral of the optical conductivity) associated with the free-carriers of the normal state (n_n) collapses into the superconducting condensate²¹ ($n_s \equiv n_n$) below T_c . Allowing that the low-frequency conductivity at $T \approx T_c$ can be described by the simple Drude theory for a metal,

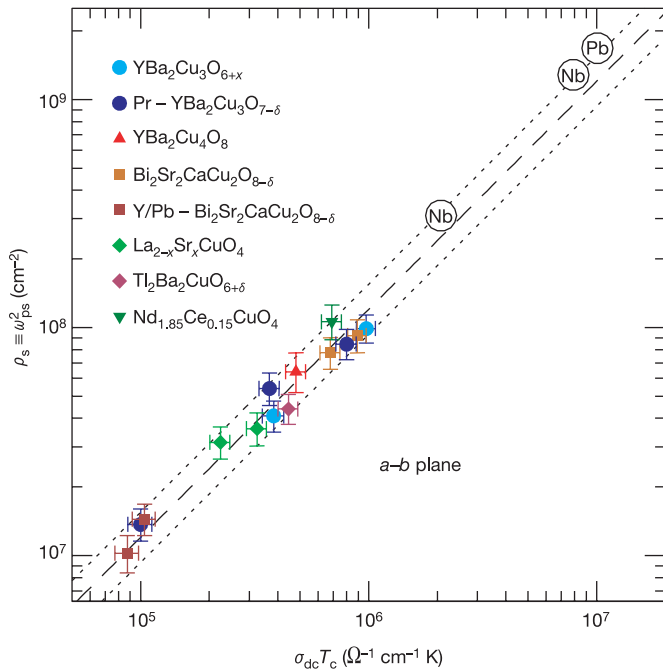


Figure 1 Plot of the superfluid density (ρ_s) versus the product of the d.c. conductivity (σ_{dc}) and the superconducting transition temperature (T_c) for a variety of copper oxides and some simple metals. (σ_{dc} is measured just above the transition, and parallel to the copper-oxygen (a - b) planes; data are shown on a log-log plot; see Supplementary Table 1 for details, including errors.) The values for σ_{dc} and ρ_s are obtained from optical measurements of the reflectance. The reflectance is a complex quantity consisting of an amplitude and a phase; in an experiment only the amplitude is usually measured. However, if the reflectance is measured over a wide frequency range, the Kramers-Kronig relation may be used to obtain the phase. Once the complex reflectance is known, then other complex optical functions may be calculated (for example, the dielectric function or the conductivity). The σ_{dc} used in this scaling relation has been extrapolated from the real part of the optical conductivity $\sigma_{dc} = \sigma_1(\omega \rightarrow 0)$ at $T \approx T_c$. For $T \ll T_c$, the response of the dielectric function to the formation of a condensate is expressed purely by the real part, $\epsilon_1(\omega) = \epsilon_\infty - \omega_{ps}^2/\omega^2$, which allows the superconducting plasma frequency ω_{ps} to be calculated from $\omega_{ps}^2 = -\omega^2 \epsilon_1(\omega)$ in the $\omega \rightarrow 0$ limit, where $\omega_{ps}^2 = 4\pi n_s e^2/m^*$ is proportional to the number of carriers in the condensate. The strength of the condensate (ρ_s) is simply $\rho_s \equiv \omega_{ps}^2$. The dashed and dotted lines are described by $\rho_s = (120 \pm 25)\sigma_{dc}T_c$. Within error, all the data for the copper oxides are described by the dashed line. The data for the conventional superconductors Nb and Pb, indicated by the atomic symbols within the circles, lie slightly above the dashed line.

$\sigma_1(\omega) = \sigma_{dc}/(1 + \omega^2\tau^2)$ (where ω is frequency), which has the shape of a lorentzian centred at zero frequency with a width at half-maximum given by the scattering rate $1/\tau$, the area under this curve may be approximated simply as σ_{dc}/τ . Transport measurements for the copper oxides²² suggest that $1/\tau$ near the transition scales linearly with T_c , so the strength of the condensate is just $\rho_s \propto \sigma_{dc}T_c$, in agreement with the observed scaling relation. This result requires that these materials approach the clean limit ($1/\tau \ll 2\Delta$, where 2Δ is the superconducting energy gap).

However, this approach cannot be applied to the properties along the c axis, because it is generally conceded that transport in this direction is incoherent, and therefore hopping rather than scattering governs the physics¹⁵. The quasi-two-dimensional nature of the copper oxides, which often includes a semiconducting or activated response of the resistivity along the c axis, has motivated the description of the superconductivity in this direction in terms of a Josephson-coupling picture^{16,17,23–26}. The c -axis penetration depth λ is then determined by the Josephson current density J_c and is $\lambda^2 = \hbar^2/8\pi d e J_c$, where $J_c = (\pi\Delta/2eR_n)\tanh(\Delta/2k_B T)$, d is the separation between the planes, and $R_n = d/\sigma_{dc}$ is the normal-state tunnelling resistance²⁴. There is convincing evidence that the energy gap in the copper oxides is d -wave in nature, containing nodes at the Fermi surface^{27,28}, making the determination of J_c difficult. However, if the coupling between the planes originates at the $(0, \pi)$, $(\pi, 0)$ points²⁹ where the gap is a maximum, Δ_0 , then we can approximate $\Delta \approx \Delta_0$. Furthermore, if $\Delta_0 \propto T_c$, then $J_c \propto T_c/R_n$ and $\rho_s \propto \sigma_{dc}T_c$, which yields the observed scaling behaviour in the c -axis direction. Despite the different nature of the transport properties parallel and perpendicular to the a - b planes, the universal scaling pertaining to both directions is an unusual and surprising result that should provide new insights into the origins of the superconductivity in these materials. □

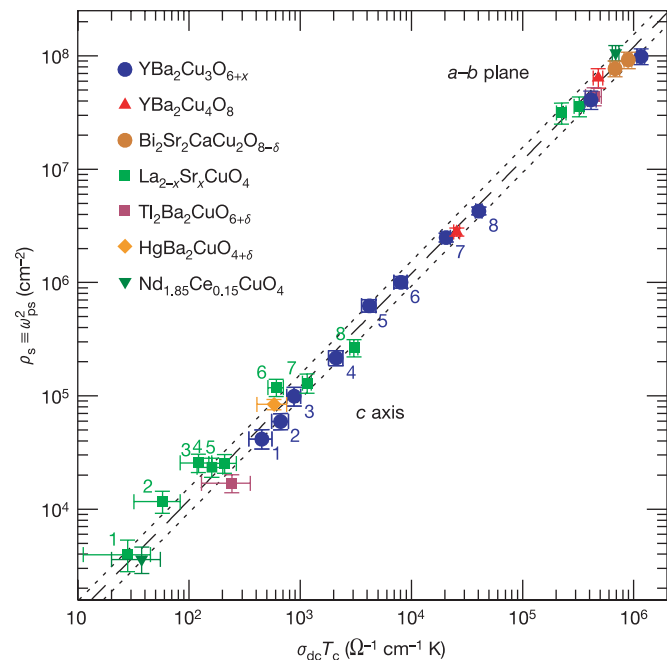


Figure 2 As Fig. 1 but for copper oxides only, and including data for the poorly conducting c axis. The values for ρ_s and σ_{dc} are obtained from optical measurements, as described in Fig. 1 legend. In addition to the published results, new data are also included for $\text{HgBa}_2\text{CuO}_{4+\delta}$ and $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$. Within error, all of the data fall on the same universal (dashed) line with slope of unity, defined by $\rho_s = 120\sigma_{dc}T_c$; the dotted lines are from $\rho_s = (120 \pm 25)\sigma_{dc}T_c$. See Supplementary Table 2 for details, including errors.

Received 20 February; accepted 17 May 2004; doi:10.1038/nature02673.

1. Bednorz, J. G. & Müller, K. A. Possible high T_c superconductivity in the Ba-La-Cu-O system. *Z. Phys. B* **64**, 189–193 (1986).
2. Uemura, Y. J. *et al.* Universal correlations between T_c and n/m^* (carrier density over effective mass) in high- T_c cuprate superconductors. *Phys. Rev. Lett.* **62**, 2317–2320 (1989).
3. Niedermayer, C. *et al.* Muon spin rotation study of the correlation between T_c and n_s/m^* in overdoped $Tl_2Ba_2CuO_{6+\delta}$. *Phys. Rev. Lett.* **71**, 1764–1767 (1993).
4. Pimenov, A. *et al.* Universal relationship between the penetration depth and the normal-state conductivity in YBCO. *Europhys. Lett.* **48**, 73–78 (1999).
5. Eisaki, H. *et al.* Effect of chemical inhomogeneity in bismuth-based copper oxide superconductors. *Phys. Rev. B* **69**, 064512 (2004).
6. Basov, D. N. *et al.* In-plane anisotropy of the penetration depth in $YBa_2Cu_3O_{7-x}$ and $YBa_2Cu_4O_8$ superconductors. *Phys. Rev. Lett.* **74**, 598–601 (1995).
7. Homes, C. C. *et al.* Effect of Ni impurities on the optical properties of $YBa_2Cu_3O_{6+x}$. *Phys. Rev. B* **60**, 9782–9792 (1999).
8. Liu, H. L. *et al.* Doping-induced change of optical properties in underdoped cuprate superconductors. *J. Phys. Condens. Matter* **11**, 239–264 (1999).
9. Puchkov, A. V., Timusk, T., Doyle, S. & Herman, A. M. *ab*-plane optical properties of $Tl_2Ba_2CuO_{6+\delta}$. *Phys. Rev. B* **51**, 3312–3315 (1995).
10. Homes, C. C., Clayman, B. P., Peng, J. L. & Greene, R. L. Optical properties of $Nd_{1.85}Ce_{0.15}CuO_4$. *Phys. Rev. B* **56**, 5525–5534 (1997).
11. Singley, E. J., Basov, D. N., Kurahashi, K., Uefuji, T. & Yamada, K. Electron dynamics in $Nd_{1.85}Ce_{0.15}CuO_{4+\delta}$: Evidence for the pseudogap state and unconventional *c*-axis response. *Phys. Rev. B* **64**, 224503 (2001).
12. Startseva, T. *et al.* Temperature evolution of the pseudogap state in the infrared response of underdoped $La_{2-x}Sr_xCuO_4$. *Phys. Rev. B* **59**, 7184–7190 (1999).
13. Pronin, A. V. *et al.* Direct observation of the superconducting energy gap developing in the conductivity spectra of niobium. *Phys. Rev. B* **57**, 14416–14421 (1998).
14. Klein, O., Nicol, E. J., Holzer, K. & Grüner, G. Conductivity coherence factors in the conventional superconductors Nb and Pb. *Phys. Rev. B* **50**, 6307–6316 (1994).
15. Ando, Y. *et al.* Metallic in-plane and divergent out-of-plane resistivity of a high- T_c cuprate in the zero temperature limit. *Phys. Rev. Lett.* **77**, 2065–2068 (1996).
16. Dordevic, S. V. *et al.* Global trends in the interplane penetration depth of layered superconductors. *Phys. Rev. B* **65**, 134511 (2002).
17. Basov, D. N., Timusk, T., Dabrowski, B. & Jorgensen, J. D. *c*-axis response of $YBa_2Cu_4O_8$: A pseudogap and possibility of Josephson coupling of CuO_2 planes. *Phys. Rev. B* **50**, 3511–3514 (1994).
18. Homes, C. C., Timusk, T., Bonn, D. A., Liang, R. & Hardy, W. N. Optical properties along the *c* axis of $YBa_2Cu_3O_{6+x}$ for $x = 0.50 \rightarrow 0.95$: Evolution of the pseudogap. *Physica C* **254**, 265–280 (1995).
19. Schützmann, J., Tajima, S., Miyamoto, S. & Tanaka, S. *c*-Axis optical response of fully oxygenated $YBa_2Cu_3O_{7-\delta}$: Observation of dirty-limit-like superconductivity and residual unpaired carriers. *Phys. Rev. Lett.* **73**, 174–177 (1994).
20. Basov, D. N. *et al.* Sum rules and interlayer conductivity of high- T_c cuprates. *Science* **283**, 49–52 (1999).
21. Tanner, D. B. *et al.* Superfluid and normal-fluid densities in high- T_c superconductors. *Physica B* **244**, 1–8 (1998).
22. Orenstein, J. *et al.* Frequency- and temperature-dependent conductivity in $YBa_2Cu_3O_{6+x}$ crystals. *Phys. Rev. B* **42**, 6342–6362 (1990).
23. Shibauchi, T. *et al.* Anisotropic penetration depth in $La_{2-x}Sr_xCuO_4$. *Phys. Rev. Lett.* **72**, 2263–2266 (1994).
24. Lawrence, W. E. & Doniach, S. in *Proc. 12th Int. Conf. Low Temperature Physics* (ed. Kando, E.) 361 (Academic, Kyoto, 1971).
25. Bulaevskii, L. N. Magnetic properties of lamellar superconductors with weak interaction between the layers. *Sov. Phys. JETP* **37**, 1133–1139 (1973).
26. Ambegaokar, V. & Baratoff, A. Tunneling between superconductors. *Phys. Rev. Lett.* **10**, 486–489 (1963).
27. Hardy, W. N., Bonn, D. A., Morgan, D. C., Liang, R. & Zhang, K. Precision measurements of the temperature dependence of λ in $YBa_2Cu_3O_{6.95}$: Strong evidence for nodes in the gap function. *Phys. Rev. Lett.* **70**, 3999–4002 (1993).
28. Shen, Z.-X. *et al.* Anomalous large gap anisotropy in the *a-b* plane of $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Phys. Rev. Lett.* **70**, 1553–1556 (1993).
29. Chakravarty, S., Sudbo, A., Anderson, P. W. & Strong, S. Interlayer tunneling and gap anisotropy in high-temperature superconductors. *Science* **261**, 337–340 (1993).

Supplementary Information accompanies this paper on www.nature.com/nature.

Acknowledgements We thank A. Chubukov, P. D. Johnson, S. A. Kivelson, P. A. Lee, D. B. Tanner, J. J. Tu, Y. Uemura and T. Valla for discussions. Work in Canada was supported by the Natural Sciences and Engineering Research Council of Canada, and the Canadian Institute for Advanced Research. The $HgBa_2CuO_{4+\delta}$ crystal growth work at Stanford University was supported by the Department of Energy's Office of Basic Energy Sciences, Division of Materials Sciences and Engineering. Work at the University of California at San Diego was supported by the National Science Foundation and the Department of Energy. Work at Brookhaven was supported by the Department of Energy.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.C.H. (homes@bnl.gov).

Magnetic phase control by an electric field

Thomas Lottermoser¹, Thomas Lonkai^{2,3}, Uwe Amann^{2,4}, Dietmar Hohlwein^{2,3}, Jörg Ihringer² & Manfred Fiebig¹

¹Max-Born-Institut, Max-Born-Straße 2A, 12489 Berlin, Germany

²Institut für Angewandte Physik, Universität Tübingen, Auf der Morgenstelle 10, 72076 Tübingen, Germany

³Hahn-Meitner-Institut, Glienicker Straße 100, 14109 Berlin, Germany

⁴Institut Laue-Langevin, 6 Rue Jules Horowitz, BP 156 - 38042 Grenoble Cedex 9, France

The quest for higher data density in information storage is motivating investigations into approaches for manipulating magnetization by means other than magnetic fields. This is evidenced by the recent boom in magnetoelectronics and 'spintronics'¹, where phenomena such as carrier effects in magnetic semiconductors² and high-correlation effects in colossal magnetoresistive compounds³ are studied for their device potential. The linear magnetoelectric effect—the induction of polarization by a magnetic field and of magnetization by an electric field—provides another route for linking magnetic and electric properties. It was recently discovered that composite materials and magnetic ferroelectrics exhibit magnetoelectric effects that exceed previously known effects^{4,5} by orders of magnitude^{6–10}, with the potential to trigger magnetic or electric phase transitions. Here we report a system whose magnetic phase can be controlled by an external electric field: ferromagnetic ordering in hexagonal $HoMnO_3$ is reversibly switched on and off by the applied field via magnetoelectric interactions. We monitor this process using magneto-optical techniques and reveal its microscopic origin by neutron and X-ray diffraction. From our results, we identify basic requirements for other candidate materials to exhibit magnetoelectric phase control.

Hexagonal $HoMnO_3$ displays ferroelectric ordering at Curie temperature $T_C = 875$ K (ref. 11), antiferromagnetic Mn^{3+} ordering at Néel temperature $T_N = 75$ K (ref. 12), and magnetic Ho^{3+} ordering at $T_{Ho} = 4.6$ K (ref. 13). The ferroelectric phase possesses $P6_3cm$ symmetry and a polarization $P_z = 5.6 \mu C cm^{-2}$ (ref. 11) along the hexagonal *z* axis. Figure 1 shows that it is made up by three magnetic sublattices with $Mn^{3+} (3d^3)$ ions at 6c positions and $Ho^{3+} (4f^{10})$ ions at 2a and 4b positions¹⁴. Anisotropy confines the Mn^{3+} spins to the basal *x-y* plane where frustration leads to four possible triangular antiferromagnetic structures^{12,15}. In contrast, the Ho^{3+} sublattices are assumed to order Ising-like along *z* showing antiferromagnetism or ferri-/ferromagnetism according to Table 1.

Magnetic Mn^{3+} ordering was monitored by optical second harmonic generation (SHG) as detailed elsewhere^{12,15}: Light at frequency ω is incident on a crystal, inducing an electromagnetic polarization at frequency 2ω , which acts as source for a SHG light wave emitted from the crystal. The magnetic symmetry determines the polarization $P(2\omega)$ of the SHG wave relative to that of the fundamental wave at ω , so that in turn $P(2\omega)$ reveals the underlying arrangement of Mn^{3+} spins. The relation between SHG polarization and Mn^{3+} ordering is tabulated elsewhere^{12,15}. Magnetic Ho^{3+} ordering was monitored by neutron powder diffraction and optical Faraday rotation, that is, the rotation $\Phi \propto \mathbf{B}$ of the plane of polarization of linearly polarized light by the magnetic field $\mathbf{B}$ in a transmission measurement. The microscopic mechanisms driving magnetoelectric phase control were revealed by neutron and X-ray powder diffraction.

Optical measurements were performed at the MBI on flux-grown, polished, *z*-oriented $HoMnO_3$ platelets ($\sim 50 \mu m$ thick) using previously described transmission set-ups^{15,16}. The static

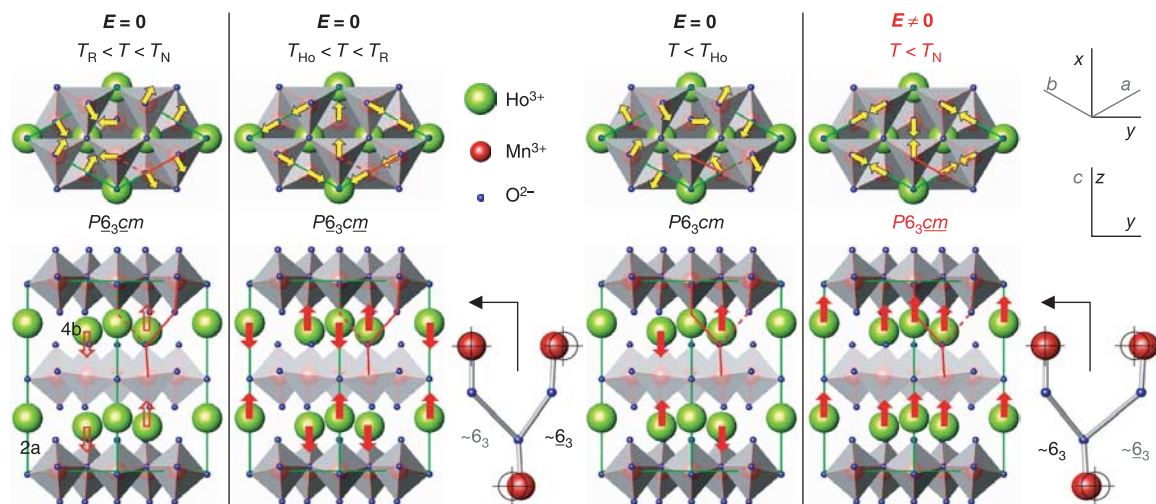


Figure 1 Magnetic structure of hexagonal HoMnO₃. Mn³⁺ and Ho³⁺ ordering in the absence and presence of an electric field $E = \pm E_0$. Arrows, magnetic moments of Mn³⁺ (yellow) and Ho³⁺ (red). Open red arrows indicate that the Ho³⁺ moments are disordered, although antiferromagnetic ordering at the 4b site is symmetry allowed. Green lines, unit cell. Red lines and Y-shaped structures, inter-plane Mn³⁺–Mn³⁺ exchange paths with

straight (dashed) lines and black (grey) labels as superior (inferior) path. Cross-wires denote the frustrated 1/3 positions of Mn³⁺ ions (see text). Deflections are enhanced by a factor of 2. Note that ferromagnetic Ho³⁺ ordering is observed exclusively when the electric field is applied.

electric field $E_0 = 10^5 \text{ V cm}^{-1}$ was applied to transparent indium-tin-oxide electrodes evaporated onto the sample surfaces. E_0 exceeds the saturation value by a factor of >10 in order to acquire a ferroelectric single-domain state even in strongly pinned samples. Neutron and X-ray experiments were performed at $E = 0$ at, respectively, the E2 beamline of the HMI and the high-resolving Guinier diffractometer at Tübingen. Neutron and X-ray diffraction data were refined simultaneously for each temperature value, using the Rietveld program SimRef¹⁷, the statistical qualifier N_σ (ref. 18), and taking the standard deviation of 2θ into account¹⁹.

Figure 2a shows the temperature dependence of SHG intensity with and without applied electric field. x -polarized fundamental light is incident along z , and the x - and y -polarized SHG contributions are detected. At $E = 0$ only x -polarized SHG light is observed below T_N , indicating $P6_3cm$ as magnetic symmetry¹⁵. It is followed at $T_R = 37 \text{ K}$ by 90° rotation of the Mn³⁺ spins and, thus, of the polarization of the SHG signal, thus changing the symmetry to $P6_3cm$ (ref. 15). At 5 K , the SHG signal is quenched owing to another 90° rotation of the Mn³⁺ spins which leads to $P6_3cm$ symmetry¹⁵. At $E = \pm E_0$, the temperature dependence of the magnetic SHG signal is markedly different. Except from a small residual surface-induced contribution²⁰, the SHG signal is quenched in the whole temperature range below T_N . The electric field forces the HoMnO₃ into a different magnetic state right at T_N , thereby controlling the magnetic order of the Mn³⁺ sublattice.

Reaction of the Ho³⁺ sublattice to the electric field is even more fundamental. The linear magnetic-field dependence of Φ (Fig. 2b) confirms the absence of any spontaneous magnetization at $E = 0$. However, Fig. 2c shows that the electric field $\pm E_0$ induces an additional contribution to the rotation. This contribution can only originate in ferromagnetic ordering of the Ho³⁺ sublattices:

It is of the same order of magnitude as the rotation induced by intrinsic ferromagnetic R^{3+} ordering in compounds with $R \in \{\text{Er, Tm, Yb}\}$ (refs. 13, 15), and it is not observed in compounds with paramagnetic R^{3+} sublattices ($R \in \{\text{Sc, Y, In, Lu}\}$). The latter point excludes additional contributions from the Mn³⁺ sublattice which remains antiferromagnetic. The electric field therefore controls the magnetic order of the Ho³⁺ system: para- or antiferromagnetism in the absence of the electric field is converted into ferromagnetic ordering with strong macroscopic magnetization (see below) in the presence of the electric field. The ferromagnetic contribution whose sign switches with reversal of the electric field is revealed by Fig. 2c, as the difference $\Delta\Phi = [\Phi(+E_0) - \Phi(-E_0)]/2$ suppresses all the ‘background’ contributions from Fig. 2b. Figure 2c shows that, in agreement with Fig. 2a, the electric-field-induced ferromagnetism persists in the whole temperature range below T_N .

A possible correlation between ferroelectricity and the existence (or non-existence²¹) of a magnetization was pointed out in experiments on BaMnF₄ (ref. 22). Further, manipulation of an existing ‘weak’ magnetization by a ferroelectric polarization was reported for Ni₃B₂O₁₃I (ref. 23). However, only in the present case the electric field creates a purely ferromagnetic Ho³⁺ state out of a para- or antiferromagnetic phase by inducing a change of magnetic structure.

Figure 3 shows the spatial distribution of Faraday rotation. The electric-field dependence of Φ in Figs 3a and b is obvious. Note that the spatial distribution of the para- and antiferromagnetic contributions and, thereby, of the sample at $E = 0$ displays variations by $>10^\circ$ while the ferromagnetic contribution $\Delta\Phi$ in Fig. 3c is homogeneous within $\pm 1^\circ$. This corroborates that the two constituents originate in different magnetic subsystems.

Table 1 reveals that the unusual example of phase control

Table 1 Magnetic structures and correlation in HoMnO₃

Magnetic symmetry	Ho (total)	Ho(2a) along z	Ho(2a) x - y plane	Ho(4b) along z	Ho(4b) x - y plane	Magnetoelectric effect
$P6_3cm$	AFM	0	0	FM	AFM	0
$P6_3cm$	AFM	AFM	FM	AFM	FM	0
$P6_3cm$	AFM	0	0	AFM	AFM	$\alpha_{xy} = -\alpha_{yx}$
$P6_3cm$	FM	FM	FM	FM	FM	$\alpha_{xx} = \alpha_{yy}, \alpha_{zz}$

Magnetic space group, allowed Ho³⁺ ordering, and magnetoelectric effect. 2a and 4b refer to the Ho³⁺ sites with ‘ x - y plane’ and ‘along z ’ referring to correlations between Ho³⁺ ions in and perpendicular to the basal plane respectively, α_{ij} denote the non-zero components of the magnetoelectric tensor. (A)FM: (anti)ferromagnetic; 0: disordered.

expressed by Figs 2 and 3 originates in the linear magnetoelectric effect. The magnetoelectric effect is described by the second-rank axial c tensor $\hat{\alpha}$ (ref. 4) and contributes a term $H_{ME} = \hat{\alpha} \mathbf{D} \mathbf{B}$ to the free energy^{5,22} (with $\mathbf{D} = \epsilon_0(\mathbf{E} + \mathbf{P})$ and $\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$ representing the sum of electric field $\mathbf{E}$ and induced polarization $\mathbf{P}$ and of magnetic field $\mathbf{H}$ and induced magnetization $\mathbf{M}$). In comparison to other compounds^{4,5}, the magnetoelectric energy H_{ME} of a magnetic ferroelectric is enhanced by orders of magnitude by the presence of ferroelectric and/or ferromagnetic ordering. When an external electric (or magnetic²⁴) field maximizes $\mathbf{D}$ (or $\mathbf{B}$) towards a single-domain state, the microscopic processes detailed below mediate the reorientation of the Mn^{3+} spins in the basal plane and the change in HoMnO_3 from antiferromagnetic to ferromagnetic interplane coupling between the Ho^{3+} spins. According to Fig. 1 and Table 1, this corresponds to a transition into the $P6_3cm$ phase where a macroscopic magnetoelectric effect from $\alpha_{zz} \neq 0$ is active. This transition costs small values of anisotropy and superexchange energy, but these are vastly overcompensated by the gain of magnetoelectric energy from $\alpha_{zz} M_z P_z$.

The microscopic processes driving the magnetoelectric effect in HoMnO_3 are revealed by Fig. 4, which shows the temperature dependence of the ordered magnetic moments of Mn^{3+} and Ho^{3+} , the position of Mn^{3+} ions, and the electric dipole moment in the unit cell as derived from neutron and X-ray powder diffraction. Figure 4a shows that even far above $T_{\text{Ho}} = 4.6$ K, long-range ordering of Ho^{3+} moments is promoted. Ho^{3+} – Mn^{3+} interaction therefore leads to an energy contribution $H_{\text{int}} = \sum \mathbf{S}^{\text{Ho}, \text{Mn}} \hat{\mathbf{A}} \mathbf{S}^{\text{Mn}}$ with $\mathbf{S}^{\text{Ho}, \text{Mn}}$ as ordered magnetic moments, the summation including all Ho^{3+} and Mn^{3+} ions in the unit cell^{24,25}. $\hat{\mathbf{A}}$ is the 3×3 interaction matrix which is made up by symmetric superexchange and asym-

metric Dzyaloshinskii-Moriya²⁶ contributions. If ferroelectric distortions are neglected, the sum is $H_{\text{int}} = 0$ for all types of Mn^{3+} ordering in Table 1. Ferroelectric distortions, however, lead to a correction $\delta \hat{\mathbf{A}} = \delta \hat{\mathbf{A}}_0 P_z$ for which—in the case of $P6_3cm$ symmetry—summation leads to the non-zero contribution $H_{\text{int}} \equiv H_{ME} = \alpha_{zz} P_z S_z^{\text{Ho}}$ with the magnetoelectric susceptibility $\alpha_{zz} = 6 S_y^{\text{Mn}} (\delta A_0^{2a} - \delta A_0^{4b})_{zy}$. Here superscripts 2a, 4b refer to the two Ho^{3+} sites (see Fig. 1).

Ho^{3+} – Mn^{3+} interaction in combination with ferroelectric

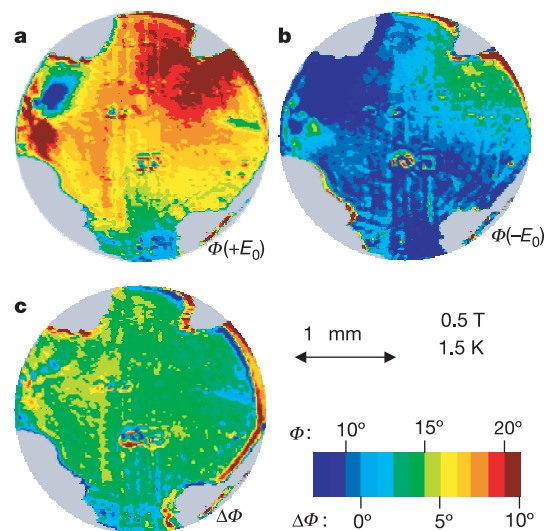


Figure 3 Spatially resolved Faraday rotation. The rotation Φ of a homogeneously polarized sample is shown as a function of the applied electric field E_0 (grey patches: electrodes). **a**, $\Phi(+E_0)$. **b**, $\Phi(-E_0)$. **c**, Contribution $\Delta\Phi = [\Phi(+E_0) - \Phi(-E_0)]/2$ induced by the applied electric field.

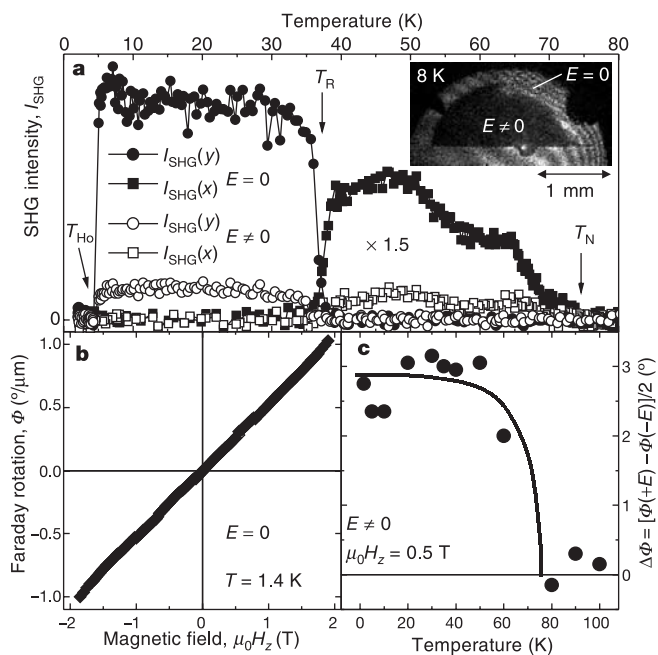


Figure 2 Dependence of magneto-optical properties of HoMnO_3 on electric field. **a**, Temperature dependence of second harmonic generation (SHG) at 2.42 eV in the absence and presence of an electric field. Filled (open) symbols refer to SHG in the absence (presence) of a field $E_0 = 10^5 \text{ V cm}^{-1}$. Circles (squares) denote the y (x) polarized components of the SHG signal. Inset, spatial distribution of SHG intensity at 2.46 eV on a sample with semicircular electrode. **b**, Faraday rotation for $E = 0$ at 1.55 eV and 1.4 K with light travelling along z , as a function of the magnetic field H_z . Para-/antiferromagnetic behaviour is observed. **c**, Temperature dependence of ferromagnetic contribution $\Delta\Phi$ to Faraday rotation for $E = \pm E_0$ at constant absorption with $\mu_0 H_z = 0.5 \text{ T}$. In $\Delta\Phi$ the ‘background’ contribution from **b** has been subtracted. The line is a guide to the eye.

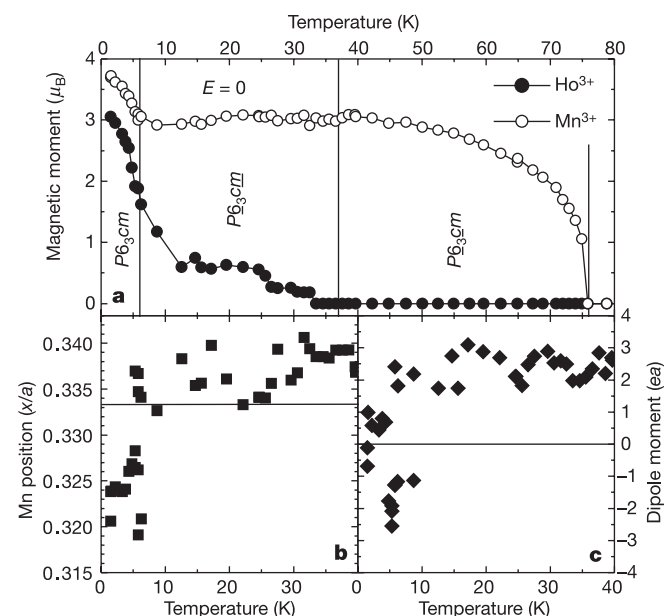


Figure 4 Microscopic manifestation of the magnetoelectric effect. Temperature dependence of **a**, ordered Mn^{3+} and Ho^{3+} moments; **b**, x position of the Mn^{3+} ions in units of in-plane lattice constant a ; **c**, ferroelectric dipole moment per unit cell. The transition of the Mn^{3+} ions from $x > a/3$ to $x < a/3$ at T_{Ho} for $E = 0$ (or, equivalently, at T_N for $E \neq 0$) reverses the competition between the Mn^{3+} – Mn^{3+} inter-plane exchange paths in Fig. 1, which leads to a conversion of the $\bar{6}_3$ into a 6_3 axis, thereby allowing the magnetoelectric effect.

distortion therefore explains the magnetoelectric coupling. Application of the electric field transfers the sample into a single-domain state with maximum field $\mathbf{D}$ for which transition into the $P6_3cm$ phase leads to the largest possible energy gain from $H_{ME} = \alpha\mathbf{D}\mathbf{B}$. According to ref. 13 and Fig. 4a up to 40% (Ho^{3+} : $3-4\mu_B$) of the rare-earth spins are ordered. The increase defining T_{Ho} in Figs 2a and 4 is due to $\text{Ho}^{3+}-\text{Ho}^{3+}$ exchange in the $x-y$ plane which complements the $\text{Ho}^{3+}-\text{Mn}^{3+}$ interaction.

Transition into the magnetoelectric $P6_3cm$ phase corresponds to modification of the inter-planar $\text{Mn}^{3+}-\text{Mn}^{3+}$ exchange paths stabilizing the three-dimensional magnetic order. With the Mn^{3+} ions at $x = \frac{1}{3}a$, the inter-planar $\text{Mn}^{3+}-\text{Mn}^{3+}$ exchange is nearly perfectly frustrated. Frustration is overcome by the $\sim 2\%$ movement of the Mn^{3+} ions revealed in Figs 1 and 4b. Depending on the shift being positive or negative, either the exchange path favouring formation of the $\underline{6}_3$ axis (above T_{Ho} at $E = 0$) or the exchange path favouring formation of the $\underline{6}_3$ axis (below T_{Ho} at $E = 0$ and below T_N at $E = \pm E_0$) is strengthened. Figure 4c shows that the electric dipole moment is modified along with the magnetic transition which is another manifestation of magnetoelectric interaction on the microscopic scale.

Thus we have observed how multifold magnetic ordering in HoMnO_3 is controlled by a static electric field. Ferromagnetic Ho^{3+} ordering is deliberately activated or deactivated, and the ferromagnetic component is controlled by the sign of the electric field. The driving mechanism for phase control are microscopic magnetoelectric interactions originating in the interplay of $\text{Ho}^{3+}-\text{Mn}^{3+}$ interaction and ferroelectric distortion. With their potential for giant magnetoelectric effects, magnetic ferroelectrics are most favourable for technological applications of magnetoelectric switching, which is reflected by the current push for novel compounds and concepts to understand this class of materials^{27,28}. On the basis of the work presented here, promising candidates for controlled magnetoelectric switching¹⁰ are compounds with electronic states close to the ground state which are energetically lowered by magnetoelectric contributions in an applied electric or magnetic field. Therefore frustrated systems or systems in the vicinity of phase boundaries or quantum critical points²⁹ are prime candidates for magnetic phase control by an electric field or vice versa. □

Received 19 April; accepted 8 June 2004; doi:10.1038/nature02728.

1. Prinz, G. A. Magnetoelectronics applications. *J. Magn. Magn. Mater.* **200**, 57–68 (1999).
2. Ohno, H. *et al.* Electric-field control of ferromagnetism. *Nature* **408**, 944–946 (2000).
3. Asamitsu, A., Tomioka, Y., Kuwahara, H. & Tokura, Y. Current switching of resistive states in magnetoresistive manganites. *Nature* **388**, 50–52 (1997).
4. O'Dell, T. H. *The Electrodynamics of Magneto-electric Media*, (North-Holland, Amsterdam, 1970).
5. Schmid, H. On a magnetoelectric classification of materials. *Int. J. Magn.* **4**, 337–361 (1973).
6. Nan, C. W. *et al.* A three-phase magnetoelectric composite of piezoelectric ceramics, rare-earth iron alloys, and polymer. *Appl. Phys. Lett.* **81**, 3831–3833 (2002).
7. Srinivasan, G., Rasmussen, E. T., Levin, B. J. & Hayes, R. Magnetoelectric effects in bilayers and multilayers of magnetoresistive and piezoelectric perovskite oxides. *Phys. Rev. B* **65**, 134402 (2002).
8. Kimura, T. Magnetic control of ferroelectric polarization. *Nature* **426**, 55–58 (2003).
9. Van Aken, B. B., Palstra, T. T. M., Filippetti, A. & Spaldin, N. A. Origin of ferroelectricity in magnetoelectric YMnO_3 . *Nature Mater.* **3**, 164–170 (2004).
10. Hur, N. *et al.* Electric polarization reversal and memory in multiferroic material induced by magnetic fields. *Nature* **429**, 392–395 (2004).
11. Coeuré, P., Guinet, F., Peuzin, J. C., Buisson, G. & Bertaut, E. F. In *Proc. Int. Meeting on Ferroelectricity* (ed. Dvorák, V.) 332–340 (Institute of Physics of the Czechoslovak Academy of Sciences, Prague, 1996).
12. Fiebig, M. *et al.* Determination of the magnetic symmetry of hexagonal manganites by second harmonic generation. *Phys. Rev. Lett.* **84**, 5620–5623 (2000).
13. Sugie, H., Iwata, N. & Kohn, K. Magnetic ordering of rare earth ions and magnetic-electric interaction of hexagonal RMnO_3 ($R = \text{Ho, Er, Yb or Lu}$). *J. Phys. Soc. Jpn* **71**, 1558–1564 (2002).
14. van Aken, B. B. *Structural Response to Electronic Transitions in Hexagonal and Ortho-Manganites*. Thesis, Reiksuniv. Groningen (2001).
15. Fiebig, M., Lottermoser, Th. & Pisarev, R. V. Spin-rotation phenomena and magnetic phase diagrams of hexagonal RMnO_3 . *J. Appl. Phys.* **93**, 8194–8197 (2003).
16. Leuter, S., Lottermoser, Th. & Fröhlich, D. Nonlinear spatially resolved phase spectroscopy. *Opt. Lett.* **24**, 1520–1522 (1999).
17. Maichle, J. K., Ihringer, J. & Prandl, W. Simultaneous structure refinement of neutron, synchrotron and x-ray powder diffraction patterns. *J. Appl. Crystallogr.* **21**, 22–28 (1998).
18. Ihringer, J. A quantitative measure for the goodness of fit in profile refinements with more than 20 degrees of freedom. *J. Appl. Crystallogr.* **28**, 618–619 (1995).

19. Rebmann, C., Ritter, H. & Ihringer, J. Standard uncertainty of angular positions and statistical quality of step-scan intensity data. *Acta Crystallogr. A* **54**, 225–231 (1998).
20. Fiebig, M., Fröhlich, D., Lottermoser, Th. & Maat, M. Probing of ferroelectric surface and bulk domains in ferroelectric RMnO_3 ($R = \text{Y, Ho}$) by second harmonic generation. *Phys. Rev. B* **66**, 144102 (2002).
21. Fox, D. L., Tilley, D. R., Scott, J. F. & Guggenheim, H. J. Magnetoelectric phenomena in BaMnF_4 and $\text{BaMn}_{0.99}\text{Co}_{0.01}\text{F}_4$. *Phys. Rev. B* **21**, 2926–2936 (1980).
22. Scott, J. F. Phase transitions in BaMnF_4 . *Rep. Prog. Phys.* **42**, 1055–1084 (1979).
23. Ascher, E., Riedel, H., Schmid, H. & Stössel, H. Some properties of ferromagnetic nickel-iodine boracite $\text{Ni}_3\text{B}_7\text{O}_{13}$. *J. Appl. Phys.* **37**, 1404–1405 (1966).
24. Fiebig, M., Degenhardt, C. & Pisarev, R. V. Interaction of frustrated magnetic sublattices in ErMnO_3 . *Phys. Rev. Lett.* **88**, 027203 (2002).
25. Kritayakirana, K., Berger, P. & Jones, R. V. Optical spectra of ferroelectric-antiferromagnetic rare earth manganates. *Opt. Commun.* **1**, 95–98 (1969).
26. Dzyaloshinskii, I. E. Thermodynamic theory of 'weak' ferromagnetism in antiferromagnetic substances. *Sov. Phys. JETP* **5**, 1259–1272 (1957).
27. Fiebig, M., Ereminko, V. & Chupis, I. (eds) *Magnetoelectric Interaction Phenomena in Crystals* (Kluwer, Dordrecht, 2004).
28. Zheng, H. *et al.* Multiferroic $\text{BaTiO}_3-\text{CoFe}_2\text{O}_4$ nanostructures. *Science* **303**, 661–663 (2004).
29. Saxena, S. S. *et al.* Superconductivity on the border of itinerant-electron ferromagnetism in UGe_2 . *Nature* **406**, 587–591 (2000).

Acknowledgements We thank K. Kohn and K. Hagdorn for samples, and the DFG and the BMBF for financial support. M. F. thanks T. Elsässer for continuous support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.F. (fiebig@mbi-berlin.de).

Demixing in simple fluids induced by electric field gradients

Yoav Tsoar, François Tournilhac & Ludwik Leibler

Laboratoire Matière Molle & Chimie (UMR 167 CNRS-ESPCI), Ecole Supérieure de Physique et Chimie Industrielles, 10 rue Vauquelin, 75231 Paris CEDEX 05, France

Phase separation in liquid mixtures is mainly controlled by temperature and pressure, but can also be influenced by gravitational, magnetic or electric fields. However, the weak coupling between such fields and concentration fluctuations limits this effect to extreme conditions^{1–3}. For example, mixing induced by uniform electric fields is detectable only at temperatures that are within a few hundredths of degree or less of the phase transition temperature of the system being studied^{4–7}. Here we predict and demonstrate that electric fields can control the phase separation behaviour of mixtures of simple liquids under more practical conditions, provided that the fields are non-uniform. By applying a voltage of 100 V across unevenly spaced electrodes about 50 μm apart, we can reversibly induce the demixing of paraffin and silicone oil at 1 K above the phase transition temperature of the mixture; when the field gradients are turned off, the mixture becomes homogeneous again. This direct control over phase separation behaviour depends on field intensity, with the electrode geometry determining the length-scale of the effect. We expect that this phenomenon will find a number of nanotechnological applications, particularly as it benefits from field gradients near small conducting objects.

The driving force for separation in liquid mixtures is the preference of constituent molecules to be in contact with their own species⁸. At high temperatures, however, thermal agitation dominates over enthalpic interactions, and mixing occurs. Figure 1a shows the classic phase diagram of a binary mixture of two liquids, A and B, with phase transition temperature, T_t (blue curve), as a function of the concentration of A (ϕ ; where $0 < \phi < 1$). Above T_t

at point 'α', the two liquids are miscible (Fig. 1b); a quench to below T_t , to a point 'β', leads to creation of (say) A-rich droplets in a B-rich continuous phase. The appearance of an interface between demixed phases is a signature of the phase separation transition (Fig. 1c).

The effect of uniform electric fields on liquid phase separation has attracted considerable attention in the past 50 years^{1,4–7,9,10}. And yet, even the question of whether the field favours mixing or demixing is still a subject of debate^{5,9,11}. Indeed, the shift of the transition temperature caused by uniform fields can originate from the non-linear dependence of dielectric constant on composition^{1,4} or from thermal composition fluctuations which in turn induce local electric field fluctuations^{10,11}. Theories predict field induced phase separation^{1,4,6,11} whereas most experiments show that uniform electric fields favour mixing^{4–7}. Still, even for mixtures with a large dielectric constant mismatch, the effect is weak: the critical temperature decreases by about 0.015 K when a field of about 4 MV m^{-1} is applied^{4,7}. However, and this is our starting point, the situation is very different when the applied field is non-uniform at a macroscopic scale. For weak fields, the mixture exhibits smooth concentration gradients, as shown schematically in Fig. 1d. This is a molecular analogue of the 'dielectric rise' effect, in which a dielectric liquid is pulled towards a region of high electric field by a dielectrophoretic force^{1,12,13}. We show that when the electric field exceeds a critical value, the composition profile changes dramatically: the mixture phase separates, creating a sharp interface (Fig. 1e). Here, demixing originates from the direct coupling between composition and field, and, therefore, the effects are strong. For simple non-polar liquids, demixing can be induced as far as a few degrees above the transition temperature and with easily accessible fields. The new displaced phase lines are shown in green and red curves in Fig. 1a, for two different field amplitudes. The electric field induced demixing may be useful for various optical and chemical applications—and once phase separation is produced, external fields can drive electrohydrodynamic¹⁴ and interfacial instabilities^{15,16} or electro-wetting phenomena^{17–19}. Upon switching off the electric field, a homogeneous mixture is recovered.

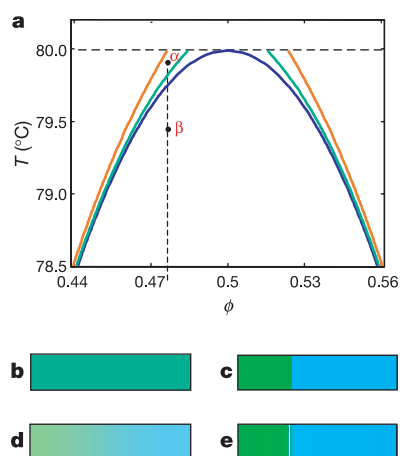


Figure 1 Phase diagram of a symmetric A/B liquid mixture. **a**, Transition temperature $T_t(\phi)$ in the absence of field (blue curve) and with a non-homogeneous field in the wedge geometry (ϕ is the fraction of A in the mixture). The green and red curves correspond to maximum fields of 8 and $15 \text{ V } \mu\text{m}^{-1}$, respectively. **b**, At point 'α' above T_t the mixture is homogeneous. **c**, At point 'β' below T_t the mixture phase separates. Green colour denotes B liquid, blue is A. **d**, Above T_t , a small field gradient induces a small concentration gradient. **e**, Above T_t , a large field gradient provokes phase separation. We used $\epsilon_A - \epsilon_B = 5$, $\partial^2 \epsilon / \partial \phi^2 = 5$, $v_0 = 3 \times 10^{-27} \text{ m}^3$ and scaled the critical temperature to be 80°C .

When an electric field is applied to a homogeneous non-conducting mixture, the electrostatic contribution of the free energy is given by F_{es} :

$$F_{\text{es}} = -\frac{1}{2} \int \epsilon(\mathbf{r}) \mathbf{E}^2(\mathbf{r}) d^3 r \quad (1)$$

where ϵ denotes the dielectric constant, and $\mathbf{E}(\mathbf{r})$ is the local field at point $\mathbf{r}$ obeying the appropriate boundary conditions on the electrodes¹. In a mixture of liquids A and B with dielectric constants ϵ_A and ϵ_B , respectively, ϵ depends on the composition of the mixture, $\phi(\mathbf{r})$. Local composition variations yield a spatially varying dielectric constant, and the system tries to adjust the mixture concentration and the local field in order to minimize the free energy.

Within a mean field approximation, the concentration profile $\phi(\mathbf{r})$ and local electric field E can be found by minimizing the free energy $F = \int f_b(\phi, T) d^3 r + F_{\text{es}}$, with boundary conditions imposed by the electrode geometry. $f_b(\phi, T)$ is the free-energy density in the absence of electric field yielding the separation temperature $T_t(\phi)$ as shown in Fig. 1 (see Methods).

For arbitrary electrode geometry analytical solution is difficult. The 'wedge' geometry, consisting of two flat and planar tilted electrodes, is particularly simple and brings a useful insight (see Fig. 2a). Indeed, for any concentration profile $\phi(r)$ with azimuthal symmetry, the electric field $\mathbf{E}$ is perpendicular to $\nabla \epsilon$ and in consequence the field $\mathbf{E}(\mathbf{r})$ is simply given by the solution of the Laplace equation $\nabla \cdot \mathbf{E} = 0$, namely, $E(r) = V/\theta r$, with V being the potential differences across the electrodes and θ the opening angle between them. This great simplification allows to find the composition profile $\phi(r)$ by solving the Euler–Lagrange equation $\delta f_b / \delta \phi - 1/2 (\delta \epsilon / \delta \phi) E^2(r) - \mu = 0$, with μ denoting the chemical potential. The analysis of this equation along classical lines²⁰ enables us to find the displacement ΔT of the transition temperature by the

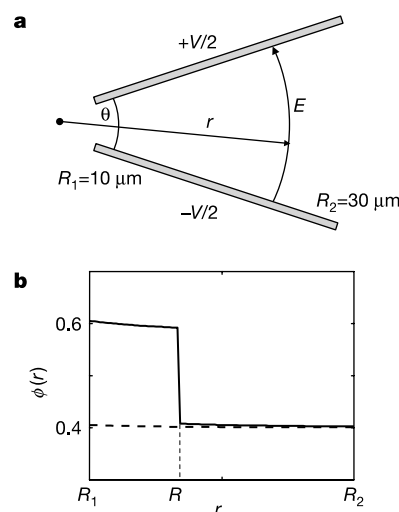


Figure 2 Wedge-shaped model system. **a**, The mixture is put between two flat electrodes with opening angle θ and potential difference V . Far from the edges, the field $E = V/\theta r$ is azimuthal. **b**, At points above the transition (for example, point 'α' of Fig. 1a) and for $\epsilon_A > \epsilon_B$, a small voltage V gives rise to a smoothly decaying profile $\phi(r)$, with high ϕ (large ϵ) at $r = R_1$ and low ϕ (small ϵ) at $r = R_2$, dashed curve. At the critical voltage V_c , the composition $\phi(R_1)$ becomes unstable. When $V > V_c$ a sharp transition is predicted between high- and low- ϕ regions, solid curve. $\phi(r)$ decays to the bulk average value at large r , here $\phi = 0.4$. We used a standard solution model for the bulk free energy f_b (see Methods section), $T = T_t + 0.2^\circ\text{C}$, $R_2 = 3R_1 = 30 \mu\text{m}$ and other parameters as in Fig. 1.

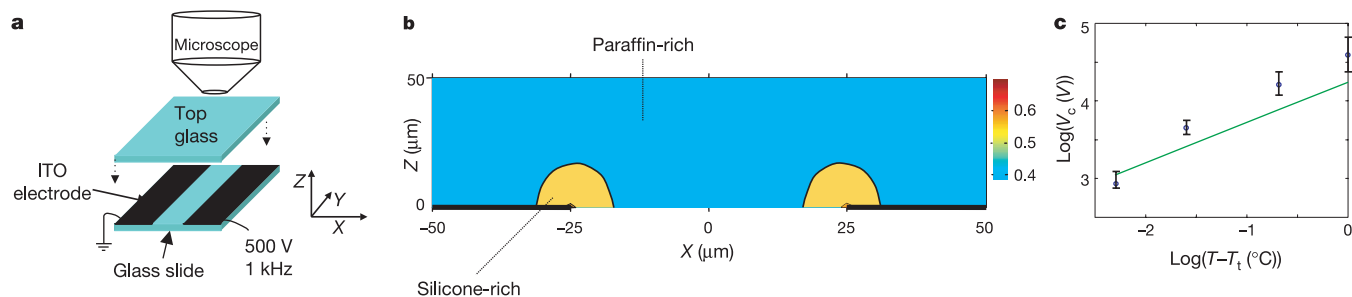


Figure 3 Phase separation with razor-blade electrodes. **a**, Two parts of the bottom glass substrate are coated with a conducting indium tin oxide layer, connected to opposite voltage terminals. The silicone/paraffin mixture between the substrate and the top glass is observed using an optical phase-contrast microscope. **b**, Predicted composition profile $\phi(r)$ in the $x-z$ plane numerically obtained for $V = 150$ V above the threshold. Black

lines indicate the two flat electrodes. The silicone-rich phase appears close to the electrode edge, while the paraffin-rich phase is farther away. Here $\phi = 0.45$, $v_0 = 3 \times 10^{-27} \text{ m}^3$ and $T - T_t = 0.2^\circ\text{C}$. **c**, Plot of $V_c(T)$ for $T > T_t$. Straight green line (slope 0.5) is model calculation, while points are experimental values with best fit slope 0.7 ± 0.15 . Error bars correspond to uncertainty in extrapolation in Fig. 4d.

above non-homogeneous field:

$$\Delta T \approx \frac{v_0}{k_B} \frac{1}{8\pi} \left| \frac{\varepsilon'}{\phi - \phi_c} \right| \left(\frac{V}{\theta R_1} \right)^2 \quad (2)$$

Here k_B is the Boltzmann constant, v_0 the liquid molecular volume, ϕ_c the critical composition and $\varepsilon' \approx \varepsilon_A - \varepsilon_B$ the derivative of the dielectric constant $\varepsilon(\phi)$ with respect to composition taken at ϕ_c . In equation (2), $r = R_1$ is the position of the inner edge of the wedge (Fig. 2a) where the field is strongest. The transition temperature is thus shifted upwards in the phase diagram (field provokes demixing) and ΔT is proportional to ε' (Fig. 1a). The estimate, equation (2), is valid for compositions ϕ such that $T_t(\phi) + \Delta T \leq T_c$. At compositions ϕ closer to ϕ_c , phase separation occurs at the critical temperature T_c (or just slightly above it when the nonlinear dependence of ε on ϕ plays some role).

The change in T_t by non-uniform fields is large even for mixtures with a weak dielectric constant mismatch, $\varepsilon_A - \varepsilon_B \approx 1$. An estimate with voltage $V = 100$ V, small feature size $R_1 = 10 \mu\text{m}$, $\theta = 1$ and $\phi = 0.4$ ($|\phi - \phi_c| \approx 0.1$) yields an upward shift of about $\Delta T \approx 0.2^\circ\text{C}$. Larger dielectric contrast ε' or smaller distance from the critical composition results in correspondingly larger ΔT .

The phase separation by a non-uniform field is much stronger than the effect of a uniform field. In a uniform field E_0 , it can be shown within the same approximations that ΔT is given by $\Delta T \approx \frac{1}{16\pi} (v_0/k_B) \varepsilon'' E_0^2$ (ref. 1). This shift is proportional to the second derivative of dielectric constant with respect to the composition ε'' and relies on second-order coupling of electric field and composition variations. Typically $\varepsilon'' \approx \varepsilon'$, and the shift ΔT in uniform field is about 10 to 50 times smaller than that in non-uniform fields, equation (2). We recall, however, that experimentally for liquid mixtures uniform field causes mixing rather than phase separation.

A useful insight into field-induced separation is gained by analysing the evolution of the composition profile for the wedge electrode geometry. When $\varepsilon' = \varepsilon_A - \varepsilon_B > 0$ and for low voltage V , the composition $\phi(R_1)$ at the inner edge is higher than $\phi(R_2)$ at the outer edge, and the profile $\phi(r)$ is a smoothly decaying function of r (Fig. 2b, dashed curve). However, at a given temperature $T \leq T_c$, there exists a threshold voltage V_c where the behaviour changes markedly. Referring to Fig. 1, the composition $\phi(R_1)$ becomes unstable when it crosses the transition curve $T_t(\phi)$, and separation occurs. An interface between A-rich (high ε) and B-rich (low ε) regions appears at $r = R$ (Fig. 2b, solid curve). The concentration profile then exhibits a sharp jump even though the field, $E(r) = V/\theta r$, still varies smoothly. When the voltage difference V increases, the interface moves towards the outer edge. From equation (2), the critical voltage V_c necessary to produce phase separation scales as $V_c \approx (T - T_t)^{1/2}$.

To test the above predictions, we used the experimental set-up

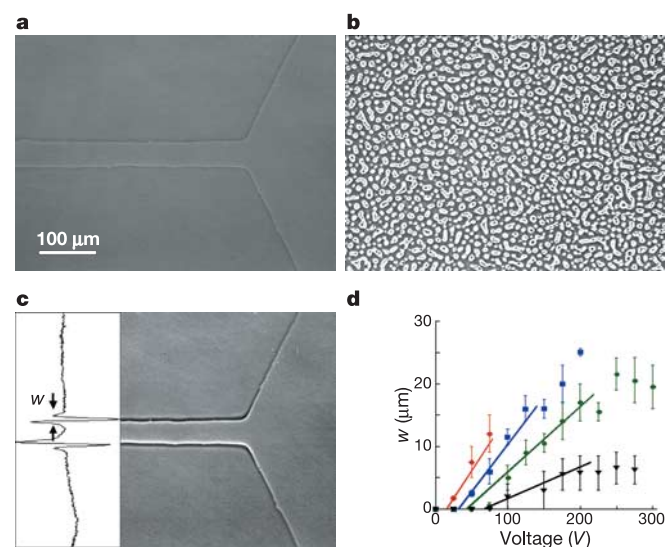


Figure 4 Temperature and voltage dependence of phase separation. **a**, The mixture is in the homogeneous state at $T = T_t + 0.1^\circ\text{C}$, and no voltage is applied. **b**, At $T = T_t - 0.3^\circ\text{C}$, isotropic phase separation occurs, in which drops begin to grow until macro-phase separation. **c**, At $T = T_t + 0.2^\circ\text{C}$ and $V = 200$ V, a silicone-rich channel appears along the electrode edges. **d**, Width w of silicone-rich phase close to the edge of the electrodes, plotted as a function of applied voltage V , for temperatures 0.1 (diamonds), 0.3 (squares), 0.5 (circles) and 1 (triangles) $^\circ\text{C}$ above T_t . Symbols are experimental data, while solid lines are results of numerical calculation (see Methods section). Error bars correspond to the scatter of three different measurements.

shown in Fig. 3a. We chose to work with razor-shaped electrodes with sharp edges in order to take advantage of the large field gradients present in such a geometry. Figure 3b shows typical concentration profiles calculated by numerical minimization of the total free energy for voltage above the threshold. Two interfaces parallel to each electrode should appear in the region close to the electrode edge. Raising the field (voltage) should displace the interfaces farther from the electrode edge, and should also increase the composition difference between coexisting phases. Lowering the temperature towards T_t while keeping the voltage fixed has a similar effect. As shown in Fig. 3c the threshold voltage V_c and the field at the electrode tips are expected to be low (below the dielectric breakdown) even for temperatures well above T_t .

We worked with two liquid pairs: a mixture of an aromatic silicone oil (polymethylphenylsiloxane) and a low-molecular-mass

paraffin oil (squalane), and a mixture of low-molecular-mass polydimethylsiloxane and polyisobutylene. We chose these systems because they are non-volatile, non-polar and have a good electric breakdown resistance. We were able to observe electric field induced demixing in both systems despite their modest dielectric constant mismatch.

Below T_b , phase separation is visible across the whole specimen (Fig. 4b). There is no preferential nucleation or wetting at the electrodes. Experiments with the electric field were performed at temperatures above T_b , namely $T - T_b = 0.1, 0.3, 0.5$ and 1°C . The influence of heating due to dielectric losses and conduction of residual ions was verified to be negligible at the working frequencies. Moreover, heating increases the temperature and opposes demixing.

Phase separation occurs immediately after the electric field is switched on, and interfaces parallel to the electrodes appear (Fig. 4c). Demixing patterns typically evolve over 1 minute, after which a channel of the silicone-rich phase (more polar liquid) is formed along the electrodes' boundaries. This nicely agrees with the numerical calculation for this electrode geometry. When the voltage is switched off, the mixture returns back to its homogeneous state in 4–5 min (Fig. 4a). The phase separation process is reversible and can be repeated many times. The universal character of field induced separation is confirmed by investigations of mixtures of polydimethylsiloxane and polyisobutylene and using different electrode spacing.

The width w of the silicone domain close to the electrode edge increases when the applied voltage is increased or when temperature is reduced closer to T_b . Figure 4d shows a plot of w as a function of V at four temperatures. There is a good agreement between the experimental points and the values obtained by numerically minimizing the free-energy functional. The theoretical and experimental critical voltages V_c , obtained from the extrapolation of each curve to the $w = 0$ axis, also agree reasonably well (Fig. 3c).

The phase separation by non-uniform electric fields is robust and, as indicated by equation (2), can be used on many liquid mixtures well below dielectric breakdown. This complements nicely the gradient field induced phase transitions in colloidal suspensions^{21,22} and ferrofluids²³. Various microfluidic or micro-electromechanical devices could benefit from this effect. For example, a mixture of several components flowing down a channel could be separated and sent off in different channels. Such a separation could provide a new way to reversibly coat the walls of the channel with a preferred chemical species and efficiently control lubrication. Chemically reactive systems could benefit as well, because active species could be isolated from each other or brought together by the electric field, thus achieving better control of reaction kinetics. Phase separation could be also employed to create reversible electro-optic effects in light guiding, scattering, and so on. In this work, non-uniform fields were created by conducting electrodes. Use of holographic optical tweezers techniques^{24,25} may open a new field of applications, such as patterning and writing. In all these applications, the reversibility of the process and the dependence on field intensity is a boon. □

Methods

Experimental

Polymethylphenylsiloxane (CAS number: 9005-12-3, Gelest PMM-0025), with a mass average molecular weight of $3,000\text{ g mol}^{-1}$ and a polydispersity index of 2.4, and squalane (CAS number: 111-01-3, purity: 99.8%) were used without further purification. Transparent indium tin oxide (ITO) electrodes 25 nm thick (surface resistivity: $120\ \Omega$ per square) were coated on the substrate glass using Shipley SJR 5740 photoresist and HCl/HNO_3 as etchant. Observation cells were made up of the ITO-treated substrate covered with a top glass layer. The temperature was regulated using a Mettler FP80 hot stage, and checked to be stable within $\pm 0.03^\circ\text{C}$ using a Pt100 probe. Before experiments, the thickness of the empty cell was measured by varying the focus of the microscope. The mixtures were heated about 20°C above their cloud point temperature before being put in the cell. The cloud point of each specimen was carefully measured through cooling cycles of $-1^\circ\text{C min}^{-1}$ and $-0.1^\circ\text{C min}^{-1}$.

The high-voltage 1 kHz a.c. source was built from a Sefram 4430 signal generator fitted

with an 18 W audio amplifier and a standard 12 V car-engine coil. Optical observations were carried out in phase contrast mode using a Leica DMRD microscope equipped with a filtered (transmission maximum at 544 nm) incandescent source and Fluotar 10×0.30 objective. Images were digitized using a JVC 3CCD camera and Eurocard Pico video board. The width w of the silicone-rich phase near the electrodes was measured from the series of light intensity fringes in the phase contrast image, as defined in the inset of Fig. 4c.

Theoretical

The numerical calculation used in Fig. 2b, Fig. 3b and c, and Fig. 4d is obtained from a variational principle of the total free energy $F = F_b + F_{es}$ with respect to the local concentration ϕ and field E :

$$\frac{\delta f_b(\phi)}{\delta \phi} - \frac{1}{2} \frac{\delta \epsilon}{\delta \phi} E^2 - \mu = 0 \quad (3a)$$

$$\nabla \cdot (\epsilon E) = 0 \quad (3b)$$

where μ is the chemical potential. The bulk free-energy density f_b is given by a Landau expansion of a standard mean-field symmetric solution model around the critical point^{20,26}: $\frac{v_0}{k_B T} f_b = \frac{1}{2} \frac{T - T_c}{T_c} (\phi - \frac{1}{2})^2 + \frac{4}{3} (\phi - \frac{1}{2})^4$, where k_B is the Boltzmann constant, v_0 is the molecular volume, T_c is the critical temperature and $\phi = 1/2$ is the critical composition. Solutions to equations (3a) and (3b) were obtained by a combination of a standard matrix inversion method for the Laplace equation and an iterative gradient scheme for the concentration profile equation.

Received 27 November 2003; accepted 14 June 2004; doi:10.1038/nature02758.

- Landau, L. D. & Lifshitz, E. M. *Elektrodinamika Sploshnykh Sred* Ch. II, Sect. 18, problem 1 (Nauka, Moscow, 1957).
- Greer, S. C., Block, T. E. & Knobler, C. M. Concentration gradients in nitroethane+3-methylpentane near the liquid-liquid critical solution point. *Phys. Rev. Lett.* **34**, 250–253 (1975).
- Moldover, M. R., Sengers, J. V., Gammon, R. V. & Hocken, R. J. Gravity effects in fluids near gas-liquid critical point. *Rev. Mod. Phys.* **51**, 79–99 (1979).
- Debye, P. & Kleboch, K. Electrical field effect on the critical opalescence. *J. Chem. Phys.* **42**, 3155–3162 (1965).
- Early, M. D. Dielectric constant measurements near the critical point of cyclohexane-aniline. *J. Chem. Phys.* **96**, 641–647 (1992).
- Wirtz, D. & Fuller, G. G. Phase transitions induced by electric fields in near-critical polymer solutions. *Phys. Rev. Lett.* **71**, 2236–2239 (1993).
- Orzechowski, K. Electric field effect on the upper critical solution temperature. *Chem. Phys.* **240**, 275–281 (1999).
- Rowlinson, J. S. & Swinton, F. S. *Liquids and Liquid Mixtures* (Butterworths, London, 1982).
- Sengers, J. V., Bedeaux, D., Mazur, P. & Greer, S. C. Behavior of the dielectric constant of fluids near a critical point. *Physica A* **104**, 573–594 (1980).
- Amundson, K., Helfand, E., Quan, X. N., Hudson, S. D. & Smith, S. D. Alignment of lamellar block-copolymer microstructure in an electric-field. 2. Mechanisms of alignment. *Macromolecules* **27**, 6559–6570 (1994).
- Onuki, A. Electric-field effects in fluids near the critical point. *Europhys. Lett.* **29**, 611–616 (1995).
- Pohl, H. A. *Dielectrophoresis* (Cambridge Univ. Press, Cambridge, UK, 1978).
- Christen, T. Homogenization of nonuniform electric fields in mixtures of liquid dielectrics. *Appl. Phys. Lett.* **76**, 230–232 (2000).
- Melcher, J. R. & Taylor, G. I. Electrohydrodynamics—A review of the role of interfacial shear stresses. *Annu. Rev. Fluid Mech.* **1**, 111–146 (1969).
- Schäffer, E., Thurn-Albrecht, T., Russell, T. P. & Steiner, U. Electrically induced structure formation and pattern transfer. *Nature* **403**, 874–877 (2000).
- Tsori, Y. & Andelman, D. Thin film diblock copolymers in electric field: Transition from perpendicular to parallel lamellae. *Macromolecules* **35**, 5161–5170 (2002).
- Berge, B. & Peseux, J. Variable focal lens controlled by an external voltage: An application of electrowetting. *Eur. Phys. J. E* **3**, 159–163 (2000).
- Comiskey, B., Albert, J. D., Yoshizawa, H. & Jacobson, J. An electrophoretic ink for all-printed reflective electronic displays. *Nature* **394**, 253–255 (1998).
- Pollack, M. G., Fair, R. B. & Shenderov, A. D. Electrowetting-based actuation of liquid droplets for microfluidic applications. *Appl. Phys. Lett.* **77**, 1725–1726 (2000).
- Lifshitz, E. M. & Landau, L. D. *Statistical Physics* 2nd edn (Butterworth-Heinemann, New York, 1980).
- Khusid, B. & Acrivos, A. Effects of interparticle electric interactions on dielectrophoresis in colloidal suspensions. *Phys. Rev. E* **54**, 5428–5435 (1996).
- Kumar, A., Qiu, Z., Acrivos, A., Khusid, B. & Jacqmin, D. Combined negative dielectrophoresis and phase separation in nondilute suspensions subject to a high-gradient ac electric field. *Phys. Rev. E* **69**, 021402 (2004).
- Rosensweig, R. E. *Ferrohydrodynamics* (Dover, New York, 1998).
- Korda, P. T., Taylor, M. B. & Grier, D. G. Kinetically locked-in colloidal transport in an array of optical tweezers. *Phys. Rev. Lett.* **89**, 128301 (2002).
- Grier, D. G. A revolution in optical manipulation. *Nature* **424**, 810–816 (2003).
- De Gennes, P. G. *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, Ithaca, New York, 1979).

Supplementary Information accompanies the paper on www.nature.com/nature

Acknowledgements We thank D. Andelman and P. G. de Gennes for discussions and suggestions, and R. Bonnecaze and A.-V. Ruzette for comments on the manuscript. We gratefully acknowledge support from ATOFINA.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.L. (Ludwik.leibler@espci.fr).

Evidence of power-law flow in the Mojave desert mantle

Andrew M. Freed¹ & Roland Bürgmann²

¹Department of Earth and Atmospheric Sciences, Purdue University, West Lafayette, Indiana 47907, USA

²Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, California 94720, USA

Studies of the Earth's response to large earthquakes can be viewed as large rock deformation experiments in which sudden stress changes induce viscous flow in the lower crust and upper mantle that lead to observable postseismic surface deformation¹. Laboratory experiments suggest that viscous flow of deforming hot lithospheric rocks is characterized by a power law in which strain rate is proportional to stress raised to a power, n (refs 2, 3). Most geodynamic models of flow in the lower crust and upper mantle, however, resort to newtonian (linear) stress-strain rate relations^{4–10}. Here we show that a power-law model of viscous flow in the mantle with $n = 3.5$ successfully explains the spatial and temporal evolution of transient surface deformation following the 1992 Landers¹¹ and 1999 Hector Mine¹² earthquakes in southern California. A power-law rheology implies that viscosity varies spatially with stress causing localization of strain, and varies temporally as stress evolves, rendering newtonian models untenable. Our findings are consistent with laboratory-derived flow law parameters for hot and wet olivine—the most abundant mineral in the upper mantle—and support the contention that, at least beneath the Mojave desert^{5,6}, the upper mantle is weaker than the lower crust.

High-temperature, high-pressure creep experiments suggest that

warm rocks deform by a thermally activated flow law of the form:

$$\dot{\epsilon} = A\sigma^n e^{(-Q/RT)} \quad (1)$$

where $\dot{\epsilon}$ is strain rate (s^{-1}), A is a pre-exponential factor ($MPa^{-n} s^{-1}$), σ is the differential stress (MPa), n is the power-law exponent, Q is the activation energy ($kJ mol^{-1}$), R is the universal gas constant ($J mol^{-1} K^{-1}$), and T is temperature (K)^{2,3}. Laboratory experiments suggest that for conditions associated with the lower crust and the upper ~200 km of the mantle¹³, flow should be accommodated by dislocation creep (involving dislocation glide and recovery processes such as dislocation climb), where the exponent, n , is usually between 2 and 4 (refs 2, 3). Dislocation creep is supported by preferred mineral-axis orientations inferred from observations of seismic anisotropy¹⁴ and similarities between microstructures observed in naturally and experimentally deformed rocks¹⁵.

The effective viscosity ($\eta = \sigma/2\dot{\epsilon}$) of a power-law material is given by:

$$\eta = \sigma^{(1-n)} e^{(Q/RT)} / 2A \quad (2)$$

The stress dependence in equation (2) implies that viscosities decrease when stresses increase, such as immediately below and after an earthquake. Pre-earthquake viscosities will be restored as earthquake-related stresses are dissipated by viscous flow. With a high power-law exponent ($n > 3$), the coseismic decrease in viscosity can be quite significant, inducing very rapid early postseismic strain rates. Compared to newtonian ($n = 1$) rheologies, the reduction of postseismic strain rates associated with power-law flow would be highly accelerated. Thus, a close monitoring of deformation rates as a function of time, such as that provided by global positioning system (GPS) time-series data, can serve as a diagnostic test to differentiate between power-law and newtonian rheologies.

A hint as to the nonlinear nature of postseismic flow was evident in previous newtonian studies of the 1992 $M_w = 7.3$ Landers and 1999 $M_w = 7.1$ Hector Mine earthquakes that occurred in the Mojave desert of southern California. In a post-Hector Mine study⁵, the inferred average viscosity for the first eight months of postseismic relaxation was an order of magnitude smaller than that inferred from deformation measurements spanning a three-year period starting three months after the Landers earthquake⁶, leading Pollitz *et al.*⁵ to infer a nonlinear rheology. Refs 16 and 7 also suggested the possibility of nonlinear rheology on the basis of observations of strain-rate change in post-Hector Mine and post-1906 San Francisco earthquake time-series data, respectively.

The Landers and Hector Mine earthquakes occurred only 20 km apart, making it likely that the postseismic observations of both events result from the response of the same lithosphere. The coupled nature of these earthquakes thus makes them ideal for a stringent rheology study in that a candidate rheologic model must satisfy the postseismic observations associated with both events. We focus here on two particular data sets that allow us to address the evolution of deformation with time: nine repeatedly surveyed campaign GPS

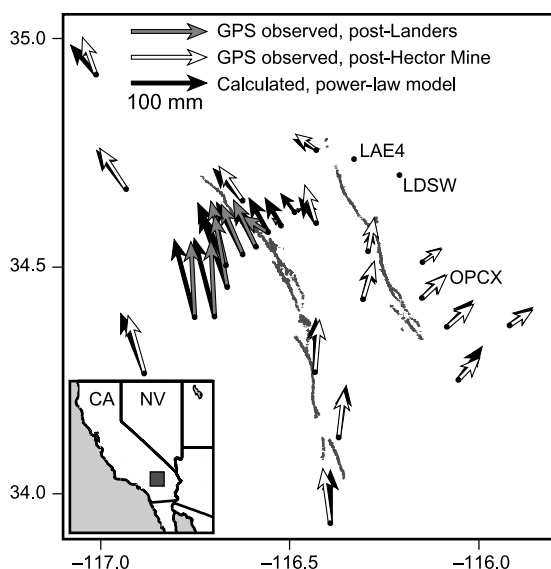


Figure 1 GPS observed and calculated horizontal postseismic surface displacements following the 1992 Landers and 1999 Hector Mine earthquakes in the Mojave desert of southern California. Two separate GPS data sets are used: campaign data following the Landers earthquake¹¹ and continuous data following the Hector Mine earthquake¹². The campaign data (grey arrows) are shown relative to station LAE4, while the continuous data (white arrows) are relative to station LDSW. Post-Landers campaign data (Emerson transect) span July 1992 through to February 1998. Post-Hector Mine continuous data span October 1999 through to December 2002. Calculated displacements are based on a finite element model that incorporates a power-law rheology of predominately mantle flow (model 8 in Fig. 2, see Table 1 for power-law parameters).

Table 1 Laboratory-derived power-law parameters considered in this analysis

Rock type	A ($MPa^{-n} s^{-1}$)	N	Q ($kJ mol^{-1}$)	Reference number
Wet quartzite1	2.2×10^{-6}	2.7	120	25
Wet quartzite2	6.3×10^{-12}	4.0	135	15
Dry quartzite 1	1.0×10^{-3}	2.0	167	26
Dry quartzite 2	3.1×10^{-4}	2.3	171	27
Aplite	3.3×10^{-7}	3.1	163	26
Quartz-diorite	1.3×10^{-3}	2.4	219	28
Anorthosite	3.2×10^{-4}	3.2	238	26
Diabase	2.2×10^{-4}	3.4	260	26
Dry olivine	1.1×10^4	3.5	535	29
Wet olivine*	3.6×10^5	3.5	480	29

For wet olivine, we fold water content into the A term by using the formulation $A^ = AC_{OH}^b$, where we assume constant $C_{OH} = 1,000H/106Si$ and use $A = 90 MPa^{-3.5} s^{-1}$ and $b = 1.2$ from ref. 29.

stations along the USGS Emerson transect¹¹ provide time series spanning both earthquakes, and 18 continuous SCIGN GPS stations provide highly precise and temporally dense measurements following the Hector Mine earthquake¹² (Fig. 1).

To infer the nature of viscous flow beneath the Mojave desert we developed a finite element model of the Landers and Hector Mine sequence that simulates coseismic slip associated with both events^{17,18}, a regional background strain rate, and rheologic models that consider both newtonian and temperature-dependent power-law rheologies. Our approach builds on our previous newtonian modelling studies of rheology and earthquake stress transfer associated with the Landers and Hector Mine sequence^{8,9}.

Previous analyses have suggested near-field (within 10–20 km) post-Landers deformation could be explained by a combination of poroelastic rebound and lower crustal flow¹⁹ or a combination of poroelastic rebound and afterslip²⁰. However, our modelling suggests that using poroelastic rebound in conjunction with viscous flow or afterslip to explain near-field post-Landers surface deformations leads to a significant underprediction of post-Landers far-field deformations (see Supplementary Fig. 1) and a pattern of post-Hector Mine surface deformations inconsistent with observations (see Supplementary Fig. 2a). In addition, the stability of the azimuth of post-Hector Mine GPS motions (see Supplementary Fig. 2b) and InSAR deformation patterns⁵ support a single

dominant deformation mechanism. We therefore consider a simpler (only one mechanism) yet advanced model that assumes power-law flow. Nevertheless, the potential contributions of multiple mechanisms cannot be ruled out.

Our analysis seeks to determine the combination of thermal gradient and assumed power laws that are consistent with the observed heat flow, geology and seismic velocities in the region, as well as with the observed postseismic deformation. We consider a range of power laws (Table 1) reflecting uncertainty in the mineralogy of the lithosphere and extrapolation from laboratory to geological conditions. Thermal gradients for the Mojave crust are estimated by assuming conductive heat transfer constrained by surface heat flow measurements that range from 52 to 69 mW m⁻², leading to inferred temperatures of 520 ± 50 to 780 ± 50 °C at the base of a 30-km-thick crust²¹. Below the crust we assume a rapid transition to an adiabatic temperature gradient (0.5 °C km⁻¹) in the shallow upper mantle based on seismic velocities that suggest a thin (5 to 10-km-thick) mantle lid overlying relatively shallow, and probably convecting, asthenosphere^{6,22}. These constraints lead us to consider a range of possible thermal gradients for the Mojave region, as shown in Fig. 2a.

Because power-law viscosity is dependent upon the absolute stress field, we begin each calculation by applying an inferred regional strain rate until steady-state stress levels are achieved in

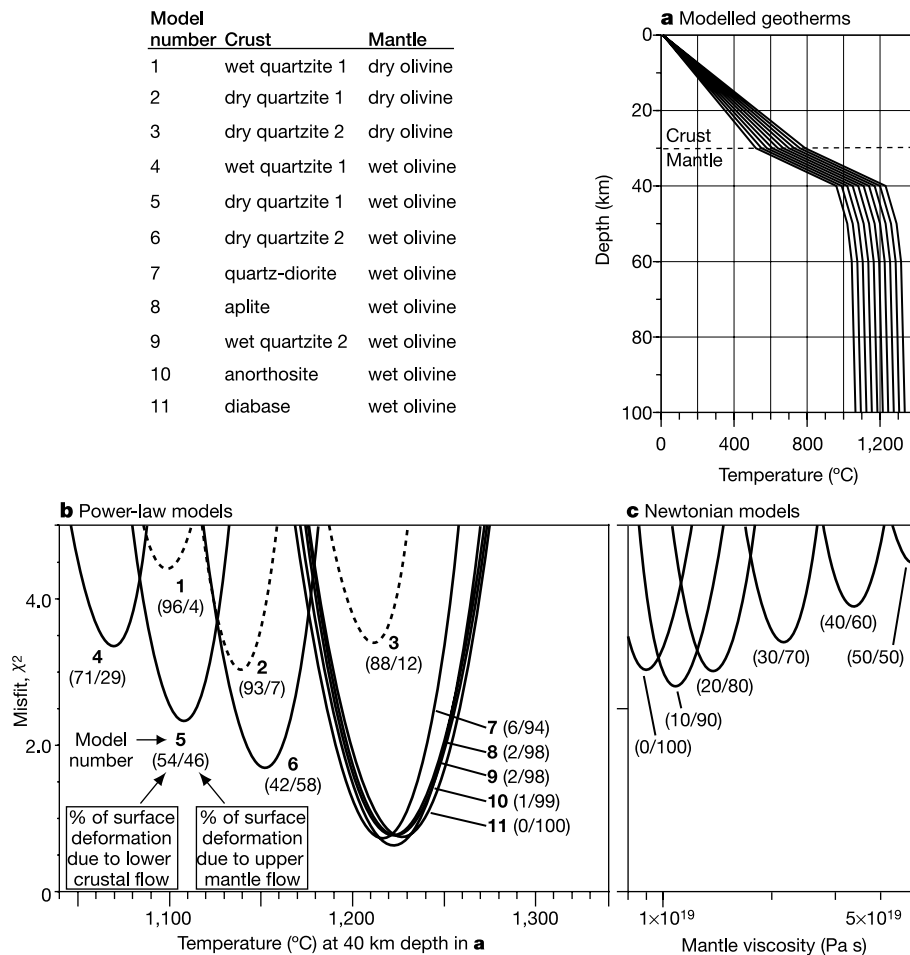


Figure 2 Testing of candidate rheologic models. Modelled geotherms (**a**) and misfits between post-Landers and post-Hector Mine GPS observed and calculated surface displacement time series for candidate power-law (**b**) and newtonian (**c**) models. Each curve in **b** represents a series of models with the same mineralogy but varying geotherm. Power-law models that consider a wet (weak) mantle are shown as solid lines, while power-law models that consider a dry (strong) mantle are shown as dotted lines. Shown

with each curve is the model number from the list above and the relative surface deformation contributed by crustal and mantle flow, respectively. Assumed power laws for each rheology are shown in Table 1. Misfits for newtonian models are shown as a function of the upper mantle viscosity (30–50 km depth). Each curve in **c** represents a newtonian model with a different relative strength between the crust and mantle, denoted by the relative percentage of surface deformation induced by flow within each layer.

regions below a depth of ~ 18 km. We assume a shear strain rate of $0.1 \times 10^{-6} \text{ yr}^{-1}$ directed along a strike of $\text{N}40^\circ \text{W}$ (ref. 11). By applying this regional strain rate to the model for several hundred years before initiating the Landers earthquake, we create a pre-Landers absolute stress state consistent with the assumed thermal gradient and rheology at depths warm enough to participate in postseismic relaxation. Once steady-state stress levels are achieved, slip associated with the 1992 Landers earthquake is applied and the crust and upper mantle are allowed to relax for 7.3 years, at which time slip associated with the 1999 Hector Mine quake is imposed. Following this, the region is allowed to relax for an additional 4 years.

In addition to power-law rheologies, we also develop a series of newtonian ($n = 1$) models in order to assess the importance of the power-law exponent in explaining the observed postseismic surface deformation. Newtonian models consider an 18-km-thick elastic upper crust, a 12-km-thick viscoelastic lower crust, and a 20-km-thick viscoelastic uppermost mantle above a 100-km-thick asthenospheric layer. A single newtonian viscosity is assigned for each layer. Because of increasing temperatures with depth, we assume that the upper mantle layer has a viscosity three times greater than the asthenosphere below.

To compare modelled and observed deformation, we define the misfit χ^2 as

$$\chi^2 = \frac{1}{N} \sum_{i=1}^N \left\{ \frac{1}{M} \sum_{j=1}^M \left[(\text{observed}_{ij} - \text{calculated}_{ij})^2 / \sigma_{ij}^2 \right] \right\} \quad (3)$$

where j are the individual time-series observations for each GPS station i , and the misfit is weighted by the estimated variance associated with each observation. We calculate misfits for the horizontal and vertical displacement time series following both earthquakes.

Figure 2b shows the misfit for various combinations of crustal and mantle power-law rheologies as a function of thermal gradient. The best models (lowest misfit) are those in which the percentage of surface deformation due to flow in the mantle (bottom number in parenthesis in Fig. 2b) is greatest. Thus, it is models that consider wet olivine in the upper mantle (solid curves in Fig. 2b) in

conjunction with a much stronger crust, such as one composed of diabase, that provide the best fit to the data. A mantle temperature of $\sim 1,300^\circ \text{C}$ at 50 km depth indicated by the best-fitting power law models is consistent with thermal models derived from seismic tomography of western North America²³ and evidence of a shallow asthenosphere inferred from Mojave desert xenoliths²⁴. Dry olivine in conjunction with a strong crust can also provide a good match to the time-series data, but the required geotherm is far greater than that allowed by heat flow constraints. In agreement with previous studies^{5,6}, lower crustal flow models lead to poor fits because the predicted vertical motions are anticorrelated with GPS observations (see Supplementary Fig. 3) and the spatial extent of horizontal surface deformation is underpredicted. The goodness of fit of a power-law model of mantle flow (model 8 in Fig. 2b, $n = 3.5$) for cumulative postseismic displacements after the Landers (5.5 yr) and Hector Mine (3.2 yr) earthquakes is shown in Fig. 1 (black arrows).

In comparison to the best power-law models, newtonian models predict the GPS observations poorly (Fig. 2c). The superior fit of the power-law model can be attributed to the manner in which the viscosity of the power-law rheology increases with time as coseismic stresses diminish (equation (2)). Although deformation rates of both newtonian and power-law rheologies diminish with time, the rate decrease is much more rapid for a power-law rheology owing to this increase in viscosity. This is illustrated in Fig. 3, which shows how two distinct newtonian models, an order of magnitude different in viscosity, are required to fit early versus late periods of the time-series data. In contrast, a single power-law rheology (solid curve) can fit the time-series curvature (Supplementary Fig. 3 shows the complete set of observed and predicted GPS time series).

The stress dependence of power-law flow inferred by our calculations means that the viscosity of the upper mantle changes as a function of time after an earthquake. Thus, a newtonian viscosity inferred from a study of surface displacements for a given period of time following a particular loading event provides only limited insight into the rheologic properties of the region, and cannot be readily extrapolated to other loading conditions, time intervals or regions. Careful integration of laboratory results with geological, geophysical and geodetic measurements promises to reveal further details of the laws governing the deformation of the Earth. $\square$

Received 14 April; accepted 24 June 2004; doi:10.1038/nature02784.

- Nur, A. & Mavko, G. Postseismic viscoelastic rebound. *Science* **183**, 204–206 (1974).
- Kirby, S. H. & Kronenberg, A. K. Rheology of the lithosphere; selected topics. *Rev. Geophys.* **25**, 1219–1244 (1987).
- Carter, N. L. & Tsenn, M. C. Flow properties of continental lithosphere. *Tectonophysics* **136**, 27–63 (1987).
- Thatcher, W., Matsuda, T., Kato, T. & Rundle, J. B. Lithospheric loading by the 1896 Riku-u earthquake, northern Japan; Implications for plate flexure and asthenospheric rheology. *J. Geophys. Res.* **85**, 6429–6435 (1980).
- Pollitz, F. F., Wicks, C. & Thatcher, W. Mantle flow beneath a continental strike-slip fault; postseismic deformation after the 1999 Hector Mine earthquake. *Science* **293**, 1814–1818 (2001).
- Pollitz, F. F., Peltzer, G. & Bürgmann, R. Mobility of continental mantle; Evidence from postseismic geodetic observations following the 1992 Landers earthquake. *J. Geophys. Res.* **105**, 8035–8054 (2000).
- Kenner, S. J. & Segall, P. Lower crustal structure in northern California: Implications from strain rate variations following the 1906 San Francisco earthquake. *J. Geophys. Res.* **108**, doi:10.1029/2001JB000189 (2003).
- Freed, A. M. & Lin, J. Delayed triggering of the 1999 Hector Mine earthquake by viscoelastic stress transfer. *Nature* **411**, 180–183 (2001).
- Freed, A. M. & Lin, J. Accelerated stress buildup on the southern San Andreas Fault and surrounding regions caused by Mojave Desert earthquakes. *Geology* **30**, 571–574 (2002).
- Deng, J., Gurnis, M., Kanamori, H. & Hauksson, E. Viscoelastic flow in the lower crust after the 1992 Landers. *Science* **282**, 1689–1692 (1998).
- Savage, J. C., Svarc, J. L. & Prescott, W. H. Near-field postseismic deformation associated with the 1992 Landers and 1999 Hector Mine, California, earthquakes. *J. Geophys. Res.* **108**, doi:10.1029/2002JB002330 (2003).
- Hudnut, K. W. et al. Continuous GPS observations of postseismic deformation following the 16 October 1999 Hector Mine, California, earthquake. *Bull. Seismol. Soc. Am.* **92**, 1403–1422 (2002).
- Karato, S., Paterson, M. S. & Fitzgerald, J. D. Rheology of synthetic olivine aggregates; influence of grain size and water. *J. Geophys. Res.* **91**, 8151–8176 (1986).
- Silver, P., Mainprice, D., Ismail, W. B. & Tommasi, A. *Mantle Petrology: Field Observations and High Pressure Experimentation: A Tribute to Francis R. (Joe) Boyd* 79–103 (Geochemical Society, Spec. Publ. No. 6, Houston, Texas, 1999).
- Hirth, G., Teyssier, C. & Dunlap, W. J. An evaluation of quartzite flow laws based on comparisons between experimentally and naturally deformed rocks. *Int. J. Earth Sci.* **90**, 77–87 (2001).

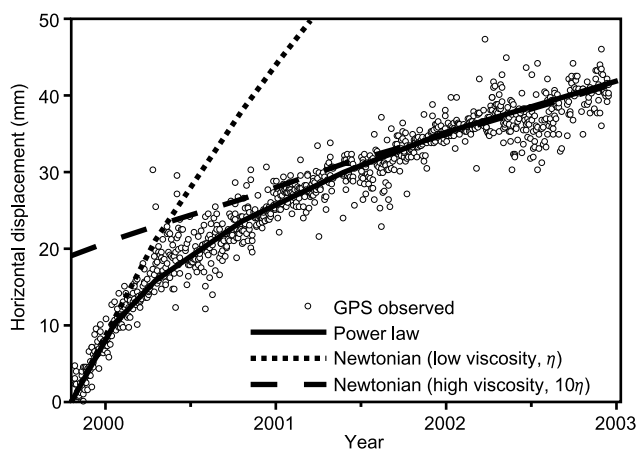


Figure 3 Comparison of representative observed and calculated postseismic displacement time series (station OPCX following the Hector Mine earthquake¹²). Power-law mantle flow model (solid curve) is model 8 in Fig. 2 (aplite, wet olivine, $7_{40 \text{ km}} = 1,225^\circ \text{C}$; power-law parameters listed in Table 1). Newtonian models consider predominately mantle flow with low viscosity ($2.5 \times 10^{18} \text{ Pa s}$, dotted curve) and an order of magnitude higher viscosity ($2.5 \times 10^{19} \text{ Pa s}$, dashed curve) that match early and late time-series slopes, respectively. The curve associated with the high-viscosity newtonian model has been raised to show where the slope matches the observed time series.

16. Pollitz, F. F. Transient rheology of the uppermost mantle beneath the Mojave Desert, California. *Earth Planet. Sci. Lett.* **215**, 89–104 (2003).
17. Wald, D. J. & Heaton, T. H. Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. *Bull. Seismol. Soc. Am.* **84**, 668–691 (1994).
18. Kaverina, A., Dreger, D. & Price, E. The combined inversion of seismic and geodetic data for the source process of the 16 October 1999 Mw 7.1 Hector Mine, California, earthquake. *Bull. Seismol. Soc. Am.* **92**, 1266–1280 (2002).
19. Masterlark, T. & Wang, H. F. Transient stress-coupling between the 1992 Landers and 1999 Hector Mine, California, earthquakes. *Bull. Seismol. Soc. Am.* **92**, 1470–1486 (2002).
20. Peltzer, G., Rosen, P. & Rogez, F. Poroelastic rebound along the Landers 1992 earthquake surface rupture. *J. Geophys. Res.* **103**, 30131–30145 (1998).
21. Williams, C. F. Temperature and the seismic/aseismic transition; observations from the 1992 Landers earthquake. *Geophys. Res. Lett.* **23**, 2029–2032 (1996).
22. Melbourne, T. & Helmberger, D. Mantle control of plate boundary deformation. *Geophys. Res. Lett.* **28**, 4003–4006 (2001).
23. Goes, S. & van der Lee, S. Thermal structure of the North American uppermost mantle inferred from seismic tomography. *J. Geophys. Res.* **107**, doi:2000JB000049 (2002).
24. Farmer, G. L. et al. Origin of late Cenozoic basalts at the Cima volcanic field, Mojave Desert, California. *J. Geophys. Res.* **100**, 8399–8415 (1995).
25. Kronenberg, A. K. & Tullis, J. A. Flow strengths of quartz aggregates; grain size and pressure effects due to hydrolytic weakening. *J. Geophys. Res.* **89**, 4281–4297 (1984).
26. Shelton, G. & Tullis, J. A. Experimental flow laws for crustal rocks. *Trans. Am. Geophys. Union* **62**, 396 (1981).
27. Jaoul, O., Tullis, J. A. & Kronenberg, A. K. The effect of varying water contents on the creep behaviour of Heavitts Quartzite. *J. Geophys. Res.* **89**, 4298–4312 (1984).
28. Hansen, F. D. & Carter, N. L. Creep of selected crustal rocks at 1000 MPa. *Trans. Am. Geophys. Union* **63**, 437 (1982).
29. Hirth, G. & Kohlstedt, D. Rheology of the upper mantle and the mantle wedge: A view from the experimentalists. In *The Subduction Factory* (ed. Eiler, J.) (American Geophysical Union 2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements Funding was provided by the National Science Foundation and the Southern California Earthquake Center. GPS data were provided by the US Geological Survey and the Southern California Integrated GPS Network.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.M.F. (freed@purdue.edu).

Fine-scale phylogenetic architecture of a complex bacterial community

Silvia G. Acinas^{1*}, Vanja Klepac-Ceraj^{1*}, Dana E. Hunt¹, Chanathip Pharino¹, Ivica Ceraj², Daniel L. Distel³ & Martin F. Polz¹

¹Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

²Bugaco, Somerville, Massachusetts 02144, USA

³Department of Biochemistry, Microbiology and Molecular Biology, University of Maine, Orono, Maine 04469, USA

* These authors contributed equally to this work

Although molecular data have revealed the vast scope of microbial diversity¹, two fundamental questions remain unanswered even for well-defined natural microbial communities: how many bacterial types co-exist, and are such types naturally organized into phylogenetically discrete units of potential ecological significance? It has been argued that without such information, the environmental function, population biology and biogeography of microorganisms cannot be rigorously explored². Here we address these questions by comprehensive sampling of two large 16S ribosomal RNA clone libraries from a coastal bacterioplankton community. We show that compensation for artefacts generated by common library construction techniques reveals fine-scale patterns of community composition. At least 516 ribotypes (unique rRNA sequences) were detected in the sample and, by statistical extrapolation, at least 1,633 co-existing

ribotypes in the sampled population. More than 50% of the ribotypes fall into discrete clusters containing less than 1% sequence divergence. This pattern cannot be accounted for by interoperon variation, indicating a large predominance of closely related taxa in this community. We propose that such micro-diverse clusters arise by selective sweeps and persist because competitive mechanisms are too weak to purge diversity from within them.

Traditional species concepts have largely been concessions to the need to identify bacteria reproducibly, but none adequately describe natural units of microbial diversity³. It has recently been proposed that natural taxa are distinct groups of strains that arise by periodic selection—a process of continuing, selectionally neutral, diversification punctuated by adaptive mutations leading to selective sweeps⁴. The latter events purge all sequence variants except those associated with the genome carrying the adaptive mutation⁴. One of the attractive features of this concept is that it should be applicable to molecular surveys of microbial diversity because taxa would be identifiable in phylogenetic trees as distinct clusters of closely related sequences¹. Moreover, such clusters should be detectable independently of the gene used to construct these trees as long as the accumulation of variation is commensurate with the occurrence of sweeps⁵. However, this theory has not been applied to broad-scale studies of bacterial diversity in the environment. Over the past 20 years, diversity studies have primarily been based on analyses of 16S rRNA clone libraries but it has remained uncertain to what extent fine-scale patterns of variation are due to sequence artefacts, to heterogeneity among paralogous operons or to the co-existence of similar but differentiated taxa¹. Furthermore, it has not been explored whether naturally defined units of differentiation emerge from recently released shotgun sequence data from the Sargasso Sea⁶.

We deduced that the discovery of ecologically significant patterns of relationships between co-existing ribotypes requires, first, an examination of clone libraries large enough to elucidate relationships at all levels of differentiation, and second, methods that minimize and account for the contribution of sequence artefacts and paralogous variation to diversity estimates. We sequenced about 1,000 clones from each of two polymerase chain reaction (PCR)-derived 16S rRNA libraries constructed from the same coastal bacterioplankton sample. The first library employed common (standard) amplification protocols. For the second, a modified protocol was designed to minimize artefacts and to identify Taq errors and chimaeric molecules through extensive sequence analyses (see Methods). This approach allowed the most comprehensive analysis of any single gene from co-occurring populations so far, even in view of the recently released Sargasso Sea study, which in aggregate sampled a similar number of rRNA genes but from several locations, dates and diverse biogeochemical conditions⁶. Our overall rationale was to achieve high coverage of rRNA genes from a single community while estimating and compensating for the influence of artefacts on ribotype diversity, potentially revealing emergent patterns.

Comparison of the two libraries showed that changes to the amplification protocol alone decreased the incidence of unique sequences from 76% (692 of 909) in the standard to 61% (686 of 1,131) in the modified library. Correction for chimaeras and Taq error lowered the percentage to 48% (516 of 1,067) unique sequences (Fig. 1a), demonstrating a potentially significant contribution of PCR-induced artefacts to (micro)diversity estimates. Consequently, these corrections yield a significantly lower estimate of total ribotype diversity for the sampled community when compared with the unmodified standard library (1,633 versus 3,881) with the use of the Chao-1 estimator⁷. A novel estimator (N_T/N_{max}) (ref. 2) yielded a similar value of 2,236 sequences for the corrected data set. This good agreement, combined with the low incidence of chimaeras and the observation that corrections account

for most expected Taq errors (see Methods), provides confidence in the corrected estimate of ribotype diversity.

A vast and previously unrecognized predominance of microdiverse ribotypes was revealed by further analysis of relationships. More than half of the observed sequences in the modified library fell into clusters sharing at least 99% sequence consensus (Fig. 1a). This result is still more marked when the Chao-1 diversity estimator is applied to the data, indicating that more than two-thirds of ribotypes might be members of 99% sequence clusters in the sampled bacterioplankton. Defining such 99% clusters, rather than unique ribotypes, as operational taxonomic units (OTUs) decreases diversity estimates from 1,633 to 520 OTUs, a decline of about 70%. However, further clustering into 98% and 97% consensus groups decreases the number of OTUs by only 3% (507 OTUs) and 11% (450 OTUs), respectively (Fig. 1a). In fact, a remarkably consistent exponential decline was observed in the number of OTUs as cluster cut-off values were decreased from 99% to 75% (Fig. 1b). In stark contrast, the number of OTUs greatly exceeds this exponential trend for values above 99% (Fig. 1b). An essentially identical relationship emerged from a phylogenetic (maximum-likelihood) analysis, in which the accumulation of lineages per arbitrary time unit was inferred under a molecular clock model⁸ (data not shown). This exponential accumulation of clusters or lineages is expected if the creation and removal of taxa are on average constant over time⁸. The sharp discontinuity observed above 99% similarity therefore suggests increased diversification or decreased removal of diversity within microdiverse clusters.

The overall predominance of extremely closely related ribotypes also emerges from phylogenetic analyses as large clusters of closely related taxa (Fig. 2, and Supplementary Information). These are typically well separated from other clusters, as indicated by a comparison of average within-cluster and between-cluster sequence divergence (data not shown)⁵. The most sequence-rich micro-

diverse clusters are formed within the *Pelagibacter* (SAR11) group of the alpha-Proteobacteria (Fig. 2) but all highly represented lineages contain such clusters, including the gamma-Proteobacteria and the Bacteroidetes group (Supplementary Information). Nevertheless, microdiverse clusters are not uniformly distributed between lineages in the modified clone library. For example, the *Cytophaga* group contains more deeply divergent lineages and fewer microdiverse clusters than the *Pelagibacter* group (Supplementary Information). However, such differences might be due to incomplete sampling because rarefaction (Fig. 1a) suggests that deeper branching lineages in this library are well sampled and that additional sequencing should therefore primarily reveal microdiverse ribotypes.

To what extent can the observed ribotype microdiversity be explained by variation among paralogous operons within single genomes¹? We have recently explored this question by an analysis of 97 available complete bacterial genomes⁹. These contain, because of multiple non-identical operons, a total of 242 different 16S rRNA sequences⁹; that is, the number of ribotypes exceeds the number of genomes about 2.5-fold. Remarkably, interoperon sequence difference remains within about 1% among these genomes. Only five genomes deviate from this rule, four of which were thermophilic bacteria all with a single operon with higher sequence divergence⁹. Therefore, if one accepts that the distribution of operons among free-living bacteria is similar to that of the 97 sequenced genomes, a conservative correction factor of about 2.5 (ref. 9) can be applied to

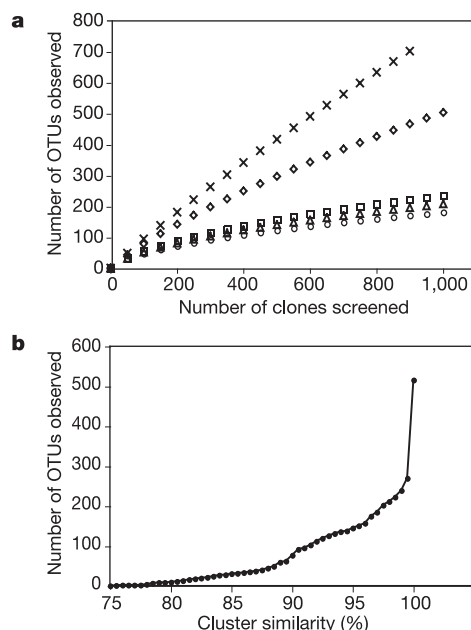


Figure 1 Compositional pattern of the coastal bacterioplankton sample. **a**, Rarefaction curves of the number of OTUs in a 16S rRNA library constructed with standard (crosses, 100% sequence similarity cluster) and modified (diamonds, 100% sequence similarity clusters; squares, 99%; triangles, 98%; circles, 97%) amplification protocols. Standard deviations fall within the symbols and are not shown. **b**, Number of OTUs plotted against changing degrees of cutoffs in 0.5% increments for grouping of sequences into similarity clusters.

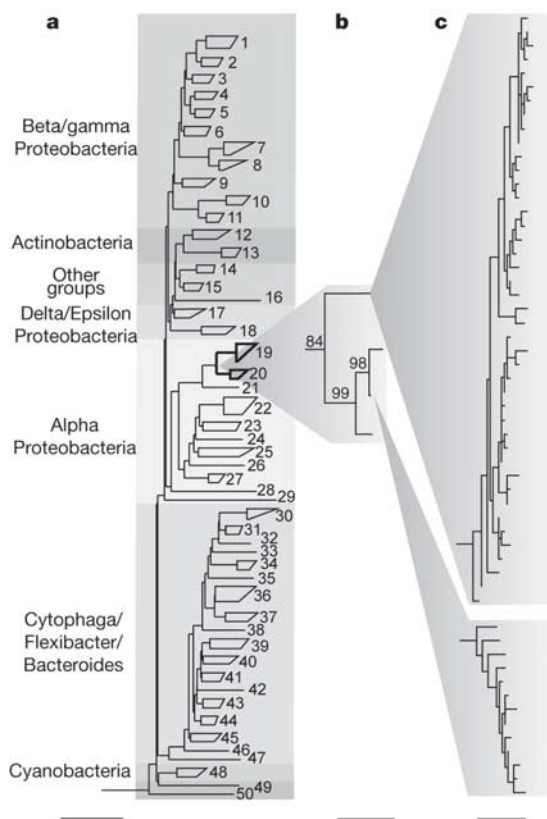


Figure 2 Phylogenetic distance relationships between the coastal bacterioplankton based on partial 16S rRNA sequencing. **a**, Summary of groups represented in the sample, in which each number denotes a phylogenetic cluster of sequences (for identification key see Supplementary Information). **b**, Relationships between *Pelagibacter* (SAR11) clusters represented by one sequence of each 99% similarity cluster. Numbers associated with nodes represent bootstrap values. **c**, Examples of microdiverse relationships between SAR11 ribotypes. Scale bars, 0.1 (**a**), 0.05 (**b**) and 0.01 (**c**) substitutions per site. Arrows connecting trees point to expanded nodes.

the estimated number of sequences (1,113) in the 99% similarity group to yield a revised estimate of at least 446 closely related genomes co-existing in the sample. However, this is probably an overcorrection because opportunistic bacteria with multiple operons are thought to predominate in culture collections and among sequenced genomes^{9,10}. Moreover, *Pelagibacter ubique* HTCC1062, which is identical in sequence to clones within the largest SAR11 99% similarity clusters, seems to contain a single rRNA operon (S. J. Giovannoni, personal communication). Given that operon numbers vary little between closely related bacteria⁹ it is unlikely that the observed SAR11 microdiversity can be explained by operon differences. Finally, shotgun sequencing of Sargasso Sea prokaryotes revealed a total of about 1,400 rRNA and about 600 RecA sequences⁶. The latter is a single-copy gene in all currently published genomes. Their frequency in the sample therefore provides an independent and almost identical estimate of 2.3 rRNA operons per genome. Thus, we conclude that, even after conservative correction, genomes denoted by microdiverse ribotypes represent by far the dominant fraction of bacterial diversity in this coastal bacterioplankton sample.

The observed pattern raises the question: what level of similarity should be expected between genomes carrying microvariant ribotypes? Comparative genomics has shown that genomes can be divided into stable and variable sets of genes, termed the core and flexible/auxiliary genome, respectively^{11,12}. The latter arises primarily by means of phage and transposon-mediated lateral gene transfer and comprises between 1% and 18% of genes¹¹ but possibly as much as 60% (ref. 13). The core genome, in contrast, is a stable complement of genes that includes rRNA and housekeeping genes. This core reflects the overall evolutionary history of the lineage because little lateral gene transfer is detectable^{9,12,14}. Microdiverse ribotype clusters should therefore also be apparent in comparisons of other housekeeping genes, possibly more so because of the higher substitution rates typical of protein coding genes⁵.

Do microdiverse sequences denote co-existing, ecologically differentiated genomes? Among free-living bacteria of very similar ribotypes, correlation of genomic variation with ecological parameters has been demonstrated convincingly in only a single case involving two strains of *Prochlorococcus*¹⁵, but these would not fall within a single microdiverse cluster as defined here. In contrast, no evidence of functional differentiation was detected in several environmental BAC clones with microdiverse 16S rRNA, despite considerable polymorphisms in protein-coding genes^{16,17}. This is consistent with recently advanced theories for the interpretation of microdiverse sequence clusters. It has been shown⁵ that clustering of housekeeping genes, resulting from periodic selection, predicts ecologically differentiated strains within cultivated bacterial taxa. If microdiverse ribotype clusters in the environment arise by the same mechanism¹ their very existence implies that intracluster competition is too weak to sweep members from within their ranks. However, this does not require that these genomes are functionally identical. Subdifferentiation within the flexible genome might provide increased fitness under episodic or spatially confined environmental conditions, but not sufficient growth advantage to sweep competing microdiverse genomes¹². Furthermore, ecological factors might decrease effective competition. Particularly, predation has been suggested to promote the coexistence of diverse lineages by 'killing the winner' of competitive events¹⁸. Finally, recombination might be important in delineating and preserving genetic diversity among members of clusters by allowing sweeps of adaptive alleles without removing selectionally neutral variation¹⁹.

The above considerations lead us to suggest that microdiverse ribotype clusters are important units of differentiation in natural bacterial communities. Indeed, such clusters might be widespread. We have recently detected numerous microdiverse ribotypes in salt-marsh sulphate-reducing bacteria²⁰, and ribotype clusters have previously been tentatively suggested for some open-ocean

microbial groups¹. To determine whether microdiverse ribotype clusters described here represent ecotypes—that is, ecologically cohesive populations—will require a detailed comparison of their encompassed genomic variation. Indeed, high-throughput sequencing⁶ and cultivation²¹ provide the means for rigorous testing of the hypothesized ecological importance of ribotype clusters. Most importantly, such inquiries challenge us to re-examine concepts of microbial diversity and invigorate the search for ecologically and evolutionarily defined species concepts. □

Methods

Study site and sampling

A 2.2-litre water sample was collected on 6 October 2001 from the marine end of the Plum Island Sound estuary (northeastern Massachusetts), and bacterioplankton was concentrated on a 0.22- μ m filter (Supor; Gelman), which was stored at -80°C until DNA extraction. Measured water parameters were as follows: temperature 16°C , pH 8.0, prokaryotic cell numbers $0.99 \times 10^9 \text{ l}^{-1}$, dissolved organic carbon 0.4 mg C l^{-1} , chlorophyll *a* $5.94 \mu\text{g l}^{-1}$.

DNA extraction and clone library construction

Cells on filters were lysed and nucleic acids were extracted with a modified version of a bead-beating method²⁰ followed by treatment with RNase I and purification on Qiagen DNA purification spin columns. Two 16S rRNA clone libraries were constructed from the same sample to estimate the total coexisting sequence diversity and the effect of PCR-induced artefacts. For both, the bacteria-specific primers 27F and 1492R as modified in ref. 22 were used. Each PCR reaction contained genomic DNA equivalent to 4.9×10^6 cells. Ten replicate reactions were combined and gel-purified, and the same amount of amplicons were cloned with the PCR 2.1-TOPO kit (Invitrogen). The PCR amplification for the first (standard) library used 35 cycles mimicking commonly used protocols (typically between 30 and 40 cycles). The second (modified) library was constructed to minimize the accumulation of the three known PCR artefacts (Taq errors, chimaeras and heteroduplex molecules)²⁰. In brief, the sample was amplified for 15 cycles followed by a 3-cycle 'reconditioning step', which eliminated heteroduplex molecules²³ and decreased the incidence of Taq errors and chimaeras. Purified plasmids served as templates for partial 16S rRNA sequence determination with the bacterial primer 27F (ref. 20).

Sequence analysis

Sequences (position 68 to 805, *E. coli* numbering) from the corrected library were further analysed for evidence of PCR artefacts. About 3% of the sequences were removed as putative chimaeras on the basis of identification by a combination of the three bioinformatics tools Chimera_Check²⁴, Bellerophon²⁵ and ChimeraBuster²⁶. Taq errors were identified by manual reconstruction of 16S rRNA secondary structures as pioneered in ref. 26 and detailed in ref. 20. In brief, the method scores as Taq errors sequence changes that violate either the sequence-conservation rule (nucleotides that are different in positions more than 98% conserved in all bacterial sequences) or the secondary-structure-conservation rule (apparent changes resulting in mismatches in stem structures that are not detected in related sequences). The Taq error rate determined from these rules was 3.3×10^{-5} per nucleotide per duplication, which agrees remarkably well with the experimentally determined value of 2×10^{-5} per nucleotide per duplication for the Taq polymerase used²⁰. Further confidence that the large majority of Taq errors are captured is lent by several simple considerations. First, inspection of alignments showed the remaining variation clustered in regions known to be highly variable, which is inconsistent with the expected random distribution of Taq errors. Second, separate quantification of Taq error rate for positions falling under the conservation and the secondary structure rule previously gave highly similar rates of 1.7×10^{-5} and 2.5×10^{-5} , respectively²⁰. Third, after 18 cycles, the inferred Taq error rate would lead to a misincorporation of bases at a rate of 3.6×10^{-4} per nucleotide. Because about 800 base pairs of sequence reads were obtained, this would translate into an average of 0.3 errors per sequence. This is close to the fraction of sequences (0.26; 181 of 686) removed from the modified library owing to identification of Taq errors. Last, potentially undetectable Taq errors by the secondary structure rule are those that change one allowed base pairing into another (for example, A-U to G-U). Although this can happen in two-thirds of all positions in rRNA, only 30% of the time will random replacement of one nucleotide by another due to Taq error result in another allowed base pairing. Therefore, at worst 20% (0.67×0.3) of Taq errors are missed but since only about 64% of the nucleotide positions fell under the secondary-structure rule this number translates into about 13%. Considering the probable incidence of errors per sequence based on a Taq error rate of 2×10^{-5} the calculation again results in an estimated 4% of sequences (0.13×0.3) that carry errors missed by the applied corrections.

The corrected set of sequences was used to estimate total diversity in the bacterioplankton community and patterns of relationships. An algorithm was developed²⁰ to group sequences into percentage similarity clusters (100%, 99%, 98%, and so on). This formed the basis for statistical extrapolation of total sequence diversity with the Chao-1 (ref. 7) and N_T/N_{MAX} (ref. 2) estimators. Accumulation of lineages through time was calculated with GENIE²⁷ from a non-optimized tree inferred by maximum likelihood with the molecular clock assumption enforced⁸. Identification of phylogenetic affiliation of the sequences was performed with the neighbour-joining method implemented in ARB²⁸ and followed by an analysis of more restricted groups of sequences by using distance and parsimony methods in PAUP* (ref. 29).

Received 10 February; accepted 11 May 2004; doi:10.1038/nature02649.

1. Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
2. Curtis, T. P., Sloan, W. T. & Scannell, J. W. Estimating prokaryotic diversity and its limits. *Proc. Natl Acad. Sci. USA* **99**, 10494–10499 (2002).
3. Rosselló-Mora, R. & Amann, R. The species concept for prokaryotes. *FEMS Microbiol. Rev.* **25**, 39–67 (2001).
4. Cohan, F. M. What are bacterial species. *Annu. Rev. Microbiol.* **56**, 457–487 (2002).
5. Palys, T., Nakamura, L. K. & Cohan, F. M. Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. *Int. J. Syst. Bacteriol.* **47**, 1145–1156 (1997).
6. Venter, C. J. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
7. Hughes, J. B., Hellmann, J. J., Ricketts, T. H. & Bohannan, B. J. M. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* **67**, 4399–4406 (2001).
8. Martin, A. P. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl. Environ. Microbiol.* **68**, 3673–3682 (2002).
9. Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J. Bacteriol.* **186**, 2629–2635 (2004).
10. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* **66**, 1328–1333 (2000).
11. Hacker, J. & Carniel, E. Ecological fitness, genomic islands and bacterial pathogenicity—a Darwinian view of the evolution of microbes. *EMBO Rep.* **2**, 376–381 (2001).
12. Lan, R. T. & Reeves, P. R. When does a clone deserve a name? A perspective on bacterial species based on population genetics. *Trends Microbiol.* **9**, 419–424 (2001).
13. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 17020–17024 (2002).
14. Daubin, V., Moran, N. A. & Ochman, H. Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829–832 (2003).
15. Rocap, G. *et al.* Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042–1047 (2003).
16. Schleper, C. A. *et al.* Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Crenarchaeum symbiosum*. *J. Bacteriol.* **180**, 5003–5009 (1998).
17. Béja, O. *et al.* Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl. Environ. Microbiol.* **68**, 335–345 (2002).
18. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45**, 1320–1328 (2000).
19. Lan, R. T. & Reeves, P. R. Intraspecific variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol.* **8**, 396–401 (2000).
20. Klepac-Ceraj, V. *et al.* High overall diversity and dominance of microdiverse relationships in salt marsh sulfate-reducing bacteria. *Environ. Microbiol.* doi: 10.1111/j.1462-2920.2004.00600.x (2004).
21. Connon, S. A. & Giovannoni, S. J. High-throughput method for culturing microorganisms in very low nutrient media yield diverse new marine isolates. *Appl. Environ. Microbiol.* **768**, 3878–3885 (2002).
22. Vergin, K. L. *et al.* Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales*. *Appl. Environ. Microbiol.* **64**, 3075–3078 (1998).
23. Thompson, J. R., Marcelino, L. A. & Polz, M. F. Heteroduplexes in mixed-template amplifications: formation, consequences and elimination by 'reconditioning PCR'. *Nucleic Acids Res.* **30**, 2083–2088 (2002).
24. Cole, J. R. *et al.* The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res.* **31**, 442–443 (2003).
25. Hugenholtz, P. & Huber, T. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int. J. Syst. Evol. Microbiol.* **53**, 289–293 (2003).
26. Field, K. G. *et al.* Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl. Environ. Microbiol.* **63**, 63–70 (1997).
27. Pybus, O. G. & Rambaut, A. GENIE: Estimating demographic history from molecular phylogenies. *Bioinformatics* **18**, 1404–1405 (2003).
28. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).
29. Swofford, D. L. *PAUP*. Phylogenetic Analysis Using Parsimony (*And Other Methods) Version 4* (Sinauer Associates, Sunderland, Massachusetts, 2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Bertilsson (Uppsala University) for sample collection, D. Veneziano (MIT) for discussions on statistical extrapolation of sequence diversity, P. Chisholm (MIT) for numerous suggestions, and researchers and staff of the PIE-LTER (Plum Island Ecosystem-Long Term Ecological Research) for logistical support. This work was supported by grants from the National Science Foundation, the Department of Energy Genomes to Life program and a postdoctoral fellowship from the Spanish Ministry of Education to S.G.A.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.F.P. (mpolz@mit.edu). The sequences of the cloned partial 16S rRNA genes have been deposited in the GenBank database under accession numbers AY580336–AY580882.

Cambrian origins and affinities of an enigmatic fossil group of arthropods

N. E. Vaccari¹, G. D. Edgecombe² & C. Escudero³

¹Instituto de Geología y Minería, Universidad Nacional de Jujuy, Avenida Bolivia 1661, San Salvador de Jujuy (4600), Argentina

²Australian Museum, 6 College Street, Sydney, NSW 2010, Australia

³O'Higgins y Zarate, 108, Palpalá (4612), Jujuy, Argentina

Euthycarcinoids are one of the most enigmatic arthropod groups, having been assigned to nearly all major clades of Arthropoda. Recent work has endorsed closest relationships with crustaceans¹ or a myriapod–hexapod assemblage², a basal position in the Euarthropoda³, or a placement in the Hexapoda⁴ or hexapod stem group⁵. Euthycarcinoids are known from 13 species ranging in age from Late Ordovician or Early Silurian to Middle Triassic, all in freshwater or brackish water environments⁶. Here we describe a euthycarcinoid from marine strata in Argentina dating from the latest Cambrian period, extending the group's record back as much as 50 million years. Despite its antiquity and marine occurrence, the Cambrian species demonstrates that morphological details were conserved in the transition to fresh water. Trackways in the same unit as the euthycarcinoid strengthen arguments that similar traces of subaerial origin from Cambro-Ordovician rocks were made by euthycarcinoids^{7,8}. Large mandibles in euthycarcinoids^{6,9} are confirmed by the Cambrian species. A morphology-based phylogeny resolves euthycarcinoids as stem-group Mandibulata, sister to the Myriapoda and Crustacea plus Hexapoda.

Mandibulata Snodgrass, 1938

Euthycarcinoidea Gall and Grauvogel, 1964

Euthycarciniformes Starobogatov, 1988

Apankura gen. nov.

Etymology. *Apankura* (Quechua), meaning crab.

Type species. *Apankura machu* gen. et sp. nov.

Diagnosis. Euthycarciniform with large mandibles that occupy most of the space beneath the posterior cephalic tergite; anterior two pairs of pre-abdominal limbs smaller than the posterior nine pairs; limbs markedly taper distally, composed of about ten podomeres, distal podomeres are shorter, large setae are absent; at least six post-abdominal segments; post-abdominal tergites are each about 2.5-times wider than they are long.

Apankura machu sp. nov.

Etymology. Genus as above; *machu* (Quechua), meaning grandfather.

Holotype. Museo de Geología, Mineralogía y Paleontología, Universidad Nacional de Jujuy (JUY-P 24; Fig. 1).

Locality and horizon. Bed of Río Huasamayo, Garganta del Diablo, near Tilcara, Jujuy Province, Argentina. The holotype (the only known specimen) is in greenish-grey mudstone from the Casa Colorada Member, Santa Rosita Formation. The trilobites *Neoparabolina frequens argentina* and *Plicatolina scalpta* on the same slab indicate a latest Cambrian age (lower part of *Neoparabolina frequens argentina* zone)¹⁰. Green shales of the Casa Colorada Member represent lower offshore deposition in an open marine facies¹¹.

Diagnosis. As for genus.

The holotype is 38 mm long, including the head, pre-abdomen and six segments of the post-abdomen. The maximum width of the pre-abdominal tergites is 16 mm. As in other euthycarcinoids^{2,12}, the head is composed of a short anterior tergite and a longer, wider posterior tergite. The latter is trapezoidal, with gently curved lateral margins. The antenna is uniramous, with at least nine short articles. Large, well-defined spheroidal processes¹³ are at the lateral margin

of the head just in front of the juncture between the cephalic tergites. These processes are transversely ovoid, about as long as the anterior cephalic tergite. A pair of large, subquadrate appendages that occupies much of the space beneath the posterior cephalic tergite is interpreted as the mandible. Dark pigmentation at their inner margins suggests enhanced sclerotization and possibly dentition on the gnathal lobe, and a palp is lacking. Immediately posterior to the mandible, patches of dark pigmentation are concentrated medially, probably corresponding to pigment strips called a buccal complex in *Euthycarcinus*⁹. In the latter, this complex resembles a setose hypopharynx. No post-mandibular oral appendages can be discerned.

The pre-abdomen has five tergites of varying length and width; the lengths increase in size from T1 to T4, with T5 about as long as T2. Their posterolateral corners appear to be blunt angulations, without spinose posterior projections. The pre-abdominal limbs are uniramous, with the posterior nine extending distally well beyond the tergal margins after flattening; anterior to these are two pairs of shorter appendages of similar form. A total of 11 pre-abdominal limbs matches other Euthycarciniformes^{9,13,14}. The limbs markedly taper distally, terminating in a spinose tip, with at least nine podomeres discerned in the best-preserved limb. The total number including the coxa may be as few as ten. The podomeres shorten distally, none of them bearing robust setae as known in some other euthycarcinoids^{6,9,13,14}. Eleven pairs of elongate, curved apodemal rods^{2,9} are preserved as dark pigment, originating near the limb

bases, and are directed posteromedially with variable degrees of rotation. Transverse anterior and posterior margins of sternites are defined across the medial third of the pre-abdomen. The lateral margins of the sternites are obscure, but given the equivalent widths of the pre-abdominal sternites and the post-abdominal segments in other euthycarcinoids, the widths of the sternites of *Apankura* are on average about 3.3 times greater than the lengths. The post-abdomen lacks appendages, and gently narrows in width posteriorly. Six segments are preserved, the width of each being about 2.5 times greater than the length. The post-abdomen is incompletely preserved posteriorly, and we infer the presence of a telson. The gut is preserved in the post-abdomen as a narrow tube with fine black fill. A band of gut filling at the posterior end of the specimen may be expelled faecal matter. A circular structure slightly offset from mid-width of the second post-abdominal segment is of uncertain identity, although a gonopore is a possibility.

Members of the Euthycarcinoidea have the following characters shared with *Apankura*: spheroidal processes at the lateral edge of the articulation between a short anterior cephalic tergite and longer posterior cephalic tergite; mandibles beneath the posterior cephalic tergite, with the mandible apparently lacking a palp; trunk tagmosis into a pedigerous pre-abdomen and a narrow, limbless post-abdomen; multiple wide sternites associated with each pre-abdominal tergite, where the ratio of sternites per tergite increases posteriorly, with decoupling of the tergal and sternal segmentation; antenniform pre-abdominal limbs with short segments; a slender, curved apodeme originating near each pre-abdominal limb base.

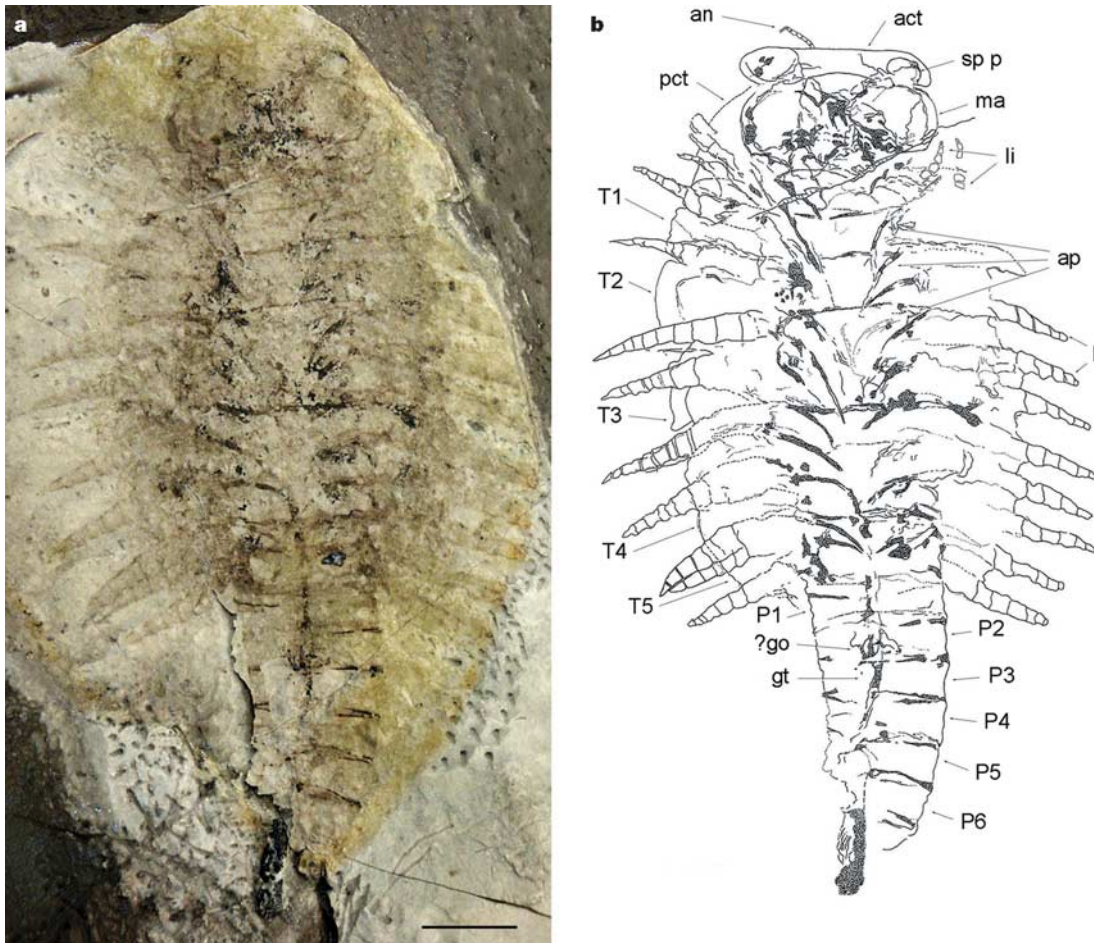


Figure 1 Holotype of *Apankura machu*. **a**, JUY-P24. Scale bar, 5 mm. **b**, Interpretative drawing of JUY-P24. act, anterior cephalic tergite; an, antenna; ap, apodemal rod; ?go, possible gonopore; gt, gut; li, limbs; ma, mandible; pct, posterior cephalic tergite;

sp p, spheroidal process; T1–T5, pre-abdominal tergites 1–5; P1–P6, post-abdominal segments 1–6.

Eleven pre-abdominal sternites, legs and apodemes and five pre-abdominal tergites are shared with other Euthycariniiformes^{9,13}. *Apankura* is unique in having reduced anterior pre-abdominal limbs, relatively few podomeres in the limbs, and at least six post-abdominal segments, versus five in the Devonian/Carboniferous Kottixeridae and four in the Carboniferous/Triassic Euthycariniidae.

The presence of mandibles as the sole gnathal appendage in euthycarinioids is indicated by the Devonian *Heterocrania*⁶ and the Carboniferous/Triassic *Euthycarcinus*^{9,14}. The mandibles of *Apankura* are of the same size as those reconstructed for *Heterocrania*⁶, and similarly occupy most of the posterior cephalic tergite (Fig. 2). A single mandibular origin in myriapods, hexapods and crustaceans (the Mandibulata) is supported by similarities in morphology and gene expression^{15,16}. Molecular data variably either support monophyly of Mandibulata¹⁷, and implicitly the homology of mandibles, or unite myriapods with chelicerates rather than with hexapods and crustaceans^{18,19}.

The phylogenetic position of euthycarinioids in the Arthropoda was evaluated by means of a matrix of 337 morphological characters (see Supplementary Information), of which euthycarinioids can be scored for 70. Analysis with equal weights resolves euthycarinioids either as a sister group to extant Mandibulata or as a sister group to Crustacea or Hexapoda within the Tetraconata (that is, Crustacea + Hexapoda)²⁰. Successive approximations²¹ and implied²² weighting select a subset of the equally weighted cladograms in which euthycarinioids form a sister group with Myriapoda and Tetraconata (Fig. 3). With this hypothesis, crown-group mandibulates are united by having the post-mandibular appendage modified as a first maxilla. The stem-group mandibulate clade that includes euthycarinioids is united by the mandible. Characters shared by myriapods and euthycarinioids, such as palpless mandibles and coxal–sternal articulation of the trunk limbs, map basally for Mandibulata. Other apomorphies of Mandibulata, such as a four-celled crystalline cone and inter-ommatidial pigment cells²³, cannot be evaluated in fossils, and whether they define the

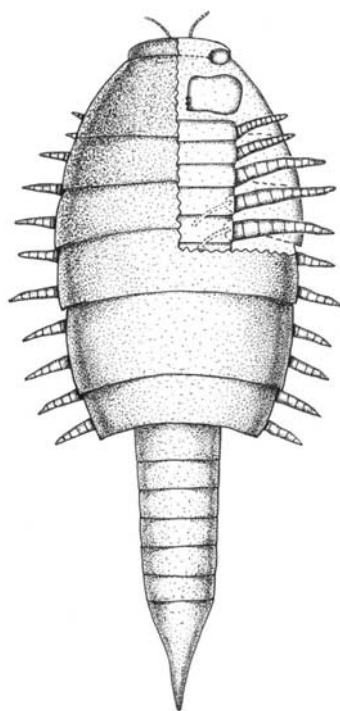


Figure 2 Reconstruction of *Apankura machu* in dorsal view, with ventral structures exposed in a cutaway view of the right side.

mandibulate crown-group or also include the stem-group is unknown.

Although *Apankura* provides by far the oldest body fossil record of euthycarinioids, trace fossils from Late Cambrian or Early Ordovician aeolian sandstones have been ascribed to an amphibious arthropod, possibly a euthycarinioid⁷. Traces similar to the “type 2” trackway⁷ attributed to euthycarinioids occur in the same unit as the holotype of *Apankura machu* (the Casa Colorada Member, at Quebrada Ruspasca; Fig. 4). The trackways are found in a mudstone and combined-flow rippled sandstone facies deposited in an upper offshore setting¹¹. The maximum preserved extent of the trackways is 12 cm long and up to 4 cm wide. At its deepest impression, the external width of the trackway is 35 mm, with its internal width²⁴

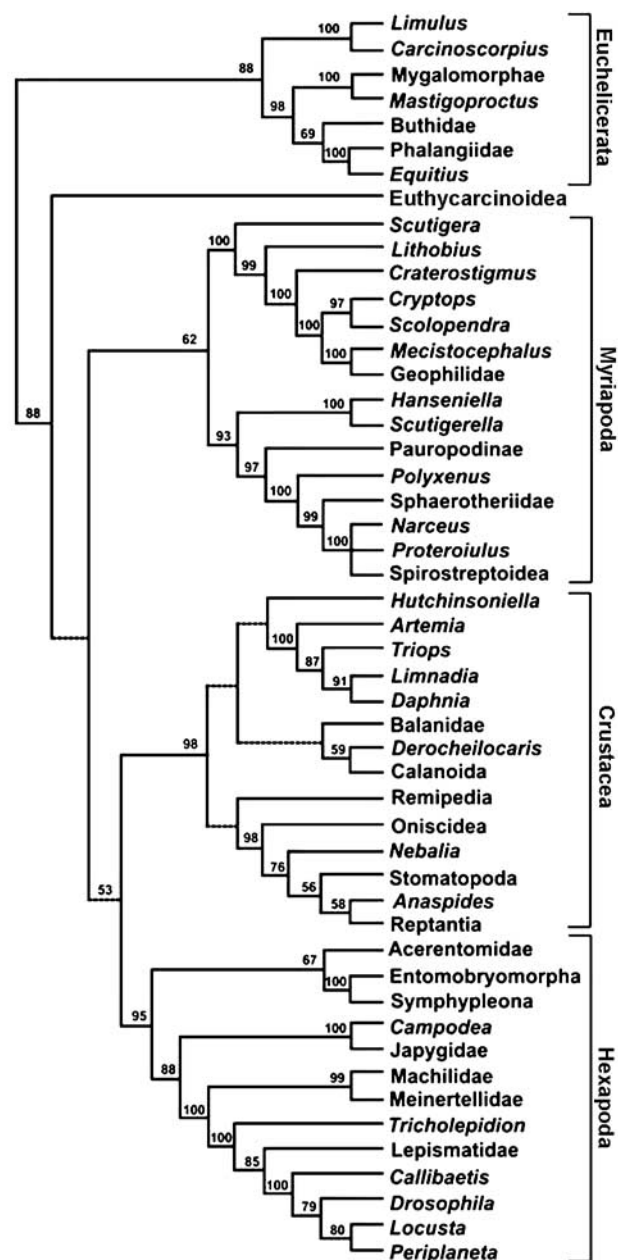


Figure 3 Strict consensus of the nine shortest cladograms for euthycarinioids and extant Euarthropoda favoured by successive weights and implied weights for concavity functions $k = 1$ to $k = 6$. Internal nodes indicated by dashed lines are collapsed in some of the 297 shortest cladograms with equal weights (673 steps, consistency index 0.63, retention index 0.86). Numbers at nodes are jack-knife frequencies above 50% with equal weights.

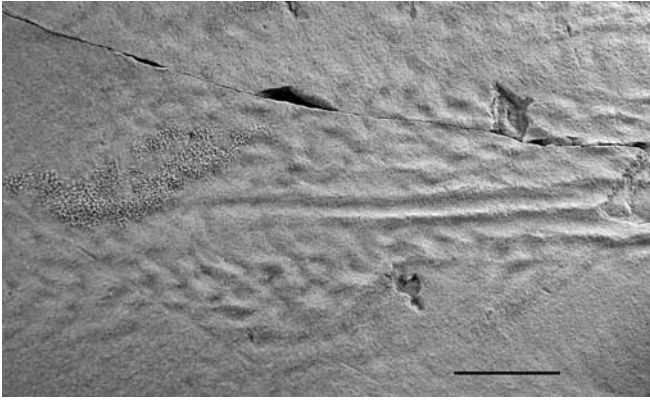


Figure 4 Trackway from the Casa Colorado Member, JUY-P 25. Scale bar, 20 mm.

being 12 mm. Imprints are obliquely elongate and too diffuse to determine an imprint series. A well-defined internal continuous mark²⁴ 5 mm wide corresponds to the telson. The presence of euthorcarinoid body fossils in the Casa Colorado Member strengthens the case of a euthorcarinoid origin for this type of trackway. Arguments that other Cambro-Ordovician trackways of possible euthorcarinoid origin were made subaerially^{7,8} signal an important role for this group in arthropod terrestrialization. □

Methods

Phylogenetic analysis

Data include 337 characters from external morphology, internal anatomy, ultrastructure, gene order and gene expression (Supplementary Information) for 52 extant arthropod terminals and the extinct clade Euthorcarinoidea. Cladograms are rooted with tardigrades and onychophorans as out-groups to Euarthropoda. A heuristic search with PAUP*4.0b10 used 10,000 random stepwise addition sequences, saving five trees per replicate, with swapping on those trees with TBR branch swapping. Multi-state taxa were scored as variable. Successive approximations weighting was implemented using the rescaled consistency indices (RCI) from the equally weighted analysis, and subjected to the same heuristic search procedures. Implied weighting explored clade sensitivity across concavity functions $k = 1-6$, with a heuristic search. Node support was assessed by 1,000 parsimony jack-knife replicates using 33% deletion.

Received 14 April; accepted 26 May 2004; doi:10.1038/nature02705.

1. Wilson, H. M. & Almond, J. E. New euthorcarinoids and an enigmatic arthropod from the British Coal Measures. *Palaeontology* **44**, 143–156 (2001).
2. Edgecombe, G. D. & Morgan, H. *Synastrus* and the euthorcarinoid puzzle. *Alcheringa* **23**, 193–213 (1999).
3. Schram, F. R. & Emerson, M. J. Arthropod Pattern Theory: a new approach to arthropod phylogeny. *Mem. Qld. Mus.* **31**, 1–18 (1991).
4. Wills, M. A., Briggs, D. E. G., Fortey, R. A., Wilkinson, M. & Sneath, P. H. A. in *Arthropod Fossils and Phylogeny* (ed. Edgecombe, G. D.) 33–105 (Columbia Univ. Press, New York, 1998).
5. McNamara, K. J. & Trewin, N. H. A euthorcarinoid arthropod from the Silurian of Western Australia. *Palaeontology* **36**, 319–335 (1993).
6. Anderson, L. I. & Trewin, N. H. An Early Devonian arthropod fauna from the Windyfield cherts, Aberdeenshire, Scotland. *Palaeontology* **46**, 467–509 (2003).
7. MacNaughton, R. B. *et al.* First steps on land: Arthropod trackways in Cambrian-Ordovician eolian sandstone, southeastern Ontario, Canada. *Geology* **30**, 391–394 (2002).
8. Trewin, N. H. & McNamara, K. J. Arthropods invade the land: Trace fossils and palaeoenvironments of the Tumbagooda Sandstone (?Late Silurian) of Kalbarri, Western Australia. *Trans. R. Soc. Edin. Earth Sci.* **85**, 117–210 (1995).
9. Gall, J.-C. & Grauvogel, L. Un arthropode peu connu. Le genre *Euthorcarinus* Handlirsch. *Ann. Paléontol. Invert.* **50**, 1–18 (1964).
10. Tortello, M. F. & Esteban, S. B. Trilobites del Cámbrico Tardío de la Formación Lampazar (sierra de Cajas, Jujuy, Argentina). Implicancias bioestratigráficas y paleoambientales. *Ameghiniana* **40**, 323–344 (2003).
11. Buatois, L. A. & Mángano, M. G. Sedimentary facies and depositional evolution of the Upper Cambrian-Lower Ordovician Santa Rosita Formation in Northwest Argentina. *J. S. Am. Earth Sci.* **16**, 343–363 (2003).
12. Bergström, J. Morphology and systematics of early arthropods. *Abh. Naturw. Ver. Hamb.* **23**, 7–42 (1980).
13. Schram, F. R. & Rolfe, W. D. I. New euthorcarinoid arthropods from the Upper Pennsylvanian of France and Illinois. *J. Paleontol.* **56**, 1434–1450 (1982).
14. Schultka, S. Erster Nachweis der Gattung *Euthorcarinus* (Arthropoda) aus dem Oberkarbon von Ibbenbüren (Nordrhein-Westfalen, Deutschland). *Paläontologische Zeitschrift* **65**, 319–332 (1991).
15. Bitsch, J. The arthropod mandible: morphology and evolution. Phylogenetic implications. *Ann. Soc. Entomol. Fr.* **37**, 305–321 (2001).
16. Edgecombe, G. D., Richter, S. & Wilson, G. D. F. The mandibular gnathal edges: homologous structures across the Mandibulata? *Afr. Invert.* **44**, 115–135 (2003).

17. Kusche, K., Hembach, A., Hagner-Holler, S., Gebauer, W. & Burmester, T. Complete subunit sequences, structure and evolution of the 6 × 6-mer hemocyanin from the common house centipede, *Scutigera coleoptrata*. *Eur. J. Biochem.* **270**, 2860–2868 (2003).
18. Mallatt, J. M., Garey, J. R. & Shultz, J. W. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol. Phylogenet. Evol.* **31**, 178–191 (2004).
19. Negrísolo, E., Minelli, A. & Valle, G. The mitochondrial genome of the house centipede *Scutigera* and the monophyly versus paraphyly of myriapods. *Mol. Biol. Evol.* **21**, 770–780 (2004).
20. Richter, S. The Tetraconata concept: hexapod–crustacean relationships and the phylogeny of Crustacea. *Org. Divers. Evol.* **2**, 217–237 (2002).
21. Farris, J. S. A successive approximations approach to character weighting. *Syst. Zool.* **18**, 374–385 (1969).
22. Goloboff, P. A. Estimating character weights during tree search. *Cladistics* **9**, 83–91 (1993).
23. Müller, C. H. G., Rosenberg, J., Richter, S. & Meyer-Rochow, V. B. The compound eye of *Scutigera coleoptrata* (Linnaeus, 1758) (Chilopoda: Notostigmophora): an ultrastructural reinvestigation that adds support to the Mandibulata concept. *Zoomorphology* **122**, 191–209 (2003).
24. Trewin, N. H. A draft system for the identification and description of arthropod trackways. *Palaeontology* **37**, 811–823 (1994).

Supplementary Information accompanies this paper on www.nature.com/nature.

Acknowledgements We thank C. Colarich, F. Rivero and E. Piovano for assistance in preparing Figs 1 and 2. Support was provided by CONICET and ANPCyT-FONCYT.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.E.V. (evac@idgym.unju.edu.ar).

Social parasitism by male-producing reproductive workers in a eusocial insect

Carlos Lopez-Vaamonde, J. Will Koning, Ruth M. Brown, William C. Jordan & Andrew F. G. Bourke

Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, UK

The evolution of extreme cooperation, as found in eusocial insects (those with a worker caste), is potentially undermined by selfish reproduction among group members^{1–3}. In some eusocial Hymenoptera (ants, bees and wasps), workers can produce male offspring from unfertilized eggs⁴. Kin selection theory predicts levels of worker reproduction as a function of the relatedness structure of the workers' natal colony and the colony-level costs of worker reproduction^{5,6}. However, the theory has been only partially successful in explaining levels of worker reproduction^{7–9}. Here we show that workers of a eusocial bumble bee (*Bombus terrestris*) enter unrelated, conspecific colonies in which they then produce adult male offspring, and that such socially parasitic workers reproduce earlier and are significantly more reproductive and aggressive than resident workers that reproduce within their own colonies. Explaining levels of worker reproduction, and hence the potential of worker selfishness to undermine the evolution of cooperation, will therefore require more than simply a consideration of the kin-selected interests of resident workers. It will also require knowledge of the full set of reproductive options available to workers, including intraspecific social parasitism.

Bombus terrestris is a common Palearctic bumble bee with an annual colony cycle, one queen per colony and single queen mating^{10–12}. Colony-level productivity costs of worker reproduction seem to be absent or low¹³. Under these conditions, kin selection theory predicts extensive male production by workers, as well as queen–worker and worker–worker conflict over worker reproduc-

Table 1 **Summary of *B. terrestris* samples genotyped at microsatellite loci**

Sample	Mean no. of individuals per colony (range)	No. of colonies	Mean no. of loci ($\pm$ s.d.) genotyped
Colony queens and mates			
Colony queens	1	30	5.9 $\pm$ 0.4
Sperm receptacles	1	5	1.2 $\pm$ 0.5
Colony queens' diploid offspring			
Workers (non-lethally sampled)	6.2 (1–10)	26	3.8 $\pm$ 1.6
Workers (died)	3.5 (1–7)	31	4.6 $\pm$ 1.4
Newly eclosed queens	2.3 (2–3)	3	6.0 $\pm$ 0.0
Newly eclosed males			
Pre-competition point in queenright colonies*	26.1 (5–90)	32	2.3 $\pm$ 1.0
Post-competition point in queenright colonies†	30.3 (1–117)	22	2.6 $\pm$ 1.1
In queenless colonies††	15.5 (1–23)	4	2.7 $\pm$ 1.9
Total males	48.8 (19–124)	32	

* Males eclosing before the competition point or within 24 days following the competition point (and hence conservatively estimated to have been produced as eggs before the competition point, because 24 days is the estimated lower limit of males' egg-to-adult developmental time²¹).

† Males eclosing in the period between 24 days after the competition point and 31 days after the queen's death, because 31 days is the estimated upper limit of male developmental time²¹.

†† Males eclosing after 31 days following the queen's death.

tion^{5,6,14}. In *B. terrestris*, consistent with this prediction¹⁵, workers lay male eggs and the onset of worker egg-laying is simultaneous with the onset of aggression between resident workers and the queen and among resident workers (at the 'competition point')^{16,17}. However, the frequency of worker-produced males has not previously been quantified, although it is believed to be low because many eggs are eaten in the fighting that follows the competition point¹⁸.

We quantified the extent of worker production of adult males in 32 semi-natural *B. terrestris* colonies using microsatellite markers (Table 1). Because we colour-marked all eclosing workers, we could classify workers within colonies as 'residents' (those present in their natal colony) and 'drifters' (those originating from other colonies, as reported in other bee species kept in aggregations of nest boxes¹⁹). Marked workers not matching their colony's colour mark (mean $\pm$ s.d. = 9.0 $\pm$ 5.1 per colony) and unmarked workers (6.0 $\pm$ 4.5 per colony) were observed within colonies (n = 32 and 29 colonies, respectively). The former were clearly drifters originating from other colonies in the 32-colony sample, but the latter could have been drifters originating from the external environment (Regent's Park, London, in which *B. terrestris* workers are common²⁰) or resident or drifter workers that had escaped marking or had lost their marks. Worker-produced males (total n = 81) were detected in 17 colonies (Table 2). In 10 colonies, resident workers were identified as the parents of worker-produced males (n = 28 males).

However, surprisingly, we found that in 13 colonies (6 of the colonies with males produced by resident workers plus 7 additional colonies; Table 2) resident workers, as well as queens, were excluded as parents of males (n = 53 males). Forty-four such males had genotypes consistent with having been produced by workers from other colonies in the 32-colony sample. Of these, 31 males were genetically attributable to workers from single additional colonies (total number of such colonies = 14), colour-marked drifter workers of which were observed in the males' natal colony in 26 cases; 13 males were genetically attributable to workers from more than one additional colony, and colour-marked drifter workers from at least one of these colonies were observed in the males' natal colony in all 13 cases. (Five males were genetically attributable to workers from single additional colonies in the 32-colony sample, but no colour-marked drifter workers from these colonies were recorded in the males' natal colonies, presumably because some drifter workers escaped observation, or had escaped marking or had lost their marks.) The other 9 worker-produced males, which were found in 5 colonies, had genotypes not attributable to any worker in the 32-colony sample, and, consistent with the presence of unmarked workers in their natal colonies, must have been produced by drifter workers from the external environment.

Queenright colonies (those whose queen had not yet died; n = 32 colonies) produced 98.5% of males, of which resident workers produced 1.9% and drifter workers produced 1.7%, with all

Table 2 **Level of worker reproduction in *B. terrestris* colonies**

Colony no.†	Number of worker-produced males detected (corrected number*)					
	Pre-competition point in queenright colonies		Post-competition point in queenright colonies		Queenless colonies	
	RW††	DW††	RW	DW	RW	DW
2	0	0	5 (6.7)	0	–§	–
3	0	0	0	0	4 (8)	6
4	0	0	1 (1.3)	1	–	–
5	0	2	0	0	0	6
6	0	0	1 (1.1)	2	–	–
11	0	0	1 (2)	0	–	–
13	0	0	1 (1.3)	0	–	–
14	0	0	0	13	–	–
15	0	2	0	0	–	–
17	0	0	0	0	8 (10.7)	4
18	0	0	0	1	–	–
19	0	0	5 (6.7)	0	–	–
23	0	1	0	0	–	–
25	0	1	0	2	–	–
26	0	1	0	0	–	–
27	1 (1.3)	0	0	0	0	1
29	0	9	1 (1.1)	1	–	–

* In the case of males produced by resident workers, the number has been corrected for non-detection error where one or more worker-produced males were detected.

† Colonies (n = 15 of 32) in which no worker-produced males were detected are omitted.

†† RW, resident worker; DW, drifter worker.

§ Blank entries (–) indicate that the colony produced no adult males when queenless.

remaining males being queen-produced ($n = 1,501$ adult males genotyped of 7,795 produced). Queenless colonies produced 1.5% of males ($n = 4$ colonies), of which resident workers produced 19.0% and drifter workers produced 28.1% ($n = 62$ adult males genotyped of 121 produced). (Males classified as having been produced by queenless colonies but genetically attributable to queens were males whose egg-to-adult developmental period must have exceeded the estimated 31-day period²¹ employed to classify males; Table 1.) At the population level, queens produced 95.7% of males, resident workers 2.2% and drifter workers 2.1%. However, in queenright colonies before the competition point, drifter workers produced significantly more males per capita than resident workers, which, as expected¹⁷ produced very few males (only one detected)(means of 1.30 and 0.00 males produced by drifter and resident workers, respectively; Wilcoxon signed rank test, $T = 27$, $n = 7$ colonies, $P < 0.05$). In queenright colonies after the competition point, there was no significant difference in per capita production of males between drifter and resident workers, although estimates were higher for the drifter workers (means of 1.01 and 0.22, respectively; Wilcoxon signed rank test, $T = 38$, $n = 10$ colonies, $P = 0.31$). In queenless colonies, the difference in per capita production of males between drifter and resident workers could not be statistically tested because too few queenless colonies produced males (Table 2), although estimates were again higher for drifter workers (means of 1.41 and 0.26, respectively). Overall, across all colonies combined, drifter workers produced significantly more males per capita than resident workers (means of 1.17 and 0.09, respectively; Wilcoxon signed rank test, $T = 130$, $n = 17$ colonies, $P = 0.01$).

Colour-marked drifter workers were observed foraging (returning with pollen loads) and, both before and after the competition point of their host colonies, egg-laying and behaving aggressively towards resident workers and other drifter workers. Drifter workers had significantly higher per capita rates of both egg-laying (means of 0.052 and 0.004 eggs per hour, respectively) and aggression (means of 0.019 and 0.002 aggressive acts per hour, respectively) than resident workers (Wilcoxon signed rank tests: egg-laying: $T = 19$, $n = 27$ colonies, $P < 0.001$; aggression: $T = 32$, $n = 21$ colonies, $P < 0.01$). Resident workers were seen killing a drifter worker as it entered the nest on one occasion, and dead drifter workers were found in the antechambers of the nests (mean $\pm$ s.d. = 3.0 ± 4.1 workers per colony), suggesting that nestmate recognition by the resident workers, although imperfect, was present. All colonies both sent and received drifter workers, but some colonies sent significantly more than they received (χ^2 test pooling colonies with low counts: $\chi^2_6 = 19.5$, $P < 0.01$), suggesting that some colonies specialized in emitting drifter workers. The number of drifter workers exported to other colonies was significantly negatively associated with the distance between the colonies' nest box exits, but was not significantly associated with the relatedness between colonies (partial Mantel tests, $n = 100,000$ permutations across 32 colonies; distance, $b = -0.33$, $P < 0.001$; relatedness, $b = -0.08$, $P = 0.08$). Therefore, drifter workers tended to come from neighbouring colonies, but did not succeed in entering colonies as a function of their relatedness to the host colony.

Although natural bumble bee nests may occur within 2 m of one another²², our colonies were at an average density higher than occurs in nature, which may have elevated the level of worker drifting. However, intraspecific social parasitism did not occur as a result of our experimental set-up, because we performed no manipulation encouraging drifter workers to reproduce and some reproductive drifter workers originated from field colonies that, moreover, must have been relatively distant (because the experimental colonies were housed on the top floor of a two-storey building). Worker drifting may occur between entirely natural *Bombus* nests, because 5–28% of workers were not genetically attributable to the resident queen in field-caught *B. hypnorum*

nests²³. A case of social parasitism similar but not identical to the present one was reported in *Vespa consobrina* wasps (workers produced males in a neighbouring *Vespa atropilosa* colony)²⁴. The present case differs from this and from the unique case of the honey bee *Apis mellifera capensis*, whose workers reproduce by thelytoky (parthenogenetic production of females) in *Apis mellifera scutellata* nests²⁵. This is because it involves a combination of intraspecific social parasitism and reproduction by the much more frequent mode, in workers of eusocial Hymenoptera⁴, of arrhenotoky (parthenogenetic production of males).

Because reproductive drifter workers did not behave like reproductive resident workers in their host or natal colonies, our findings suggest that male production by intraspecific social parasitism is a distinct reproductive tactic in *B. terrestris* workers. Drifter workers may lay eggs before their host colony's competition point because they are not under kin selection to refrain from activities reducing the production of unrelated host sexuals and because their sons would have greater mating success than males produced later^{15,26}. Whether intraspecific social parasitism occurs opportunistically in 'lost' workers or follows a deliberate emigration remains to be determined. Intraspecific social parasitism by male-producing workers may occur in other eusocial Hymenoptera, in which it could have been overlooked because detecting it requires identification of both drifter workers and their offspring. Our findings also show that resident *B. terrestris* workers successfully produce adult males in queenright colonies as kin selection theory predicts^{5,6,14}. However, the overall level of worker male-production was low, matching the expectation based on high levels of egg-eating¹⁸. Genetic studies of other bumble bee species have estimated that resident workers in queenright colonies produce 0–20% of adult males^{23,27,28}. Some of the unexplained variation in the distribution and extent of worker reproduction in bumble bees and other taxa could be owing to the capacity of workers to act as intraspecific social parasites. □

Methods

Collection and rearing of bees

We reared 32 *B. terrestris* colonies from 122 wild queens captured in Regent's Park, London and Silwood Park, Ascot, UK, in February and March 2003. When surviving queens (30 from Regent's Park and 2 from Silwood Park) had raised 10 workers, we transferred colonies to wooden nest boxes with clear Perspex lids housed in unlit rooms on the top floor of a building about 11 m high in Regent's Park. Each nest box was connected by means of an empty, identical box (antechamber) and a plastic tube (150–200 cm long $\times$ 3 cm diameter) to an exit hole in a wooden panel occupying the window frame of each room, so permitting workers to forage freely in surrounding parkland and gardens. We connected nest boxes to the exterior between 2 May and 20 June 2003 (mean $\pm$ s.d. size of colonies when connected was 36.7 ± 7.7 workers). Twenty nest boxes exited on the building's south side and twelve exited 20 m away on its north side; within each side, adjacent exit holes were 70–100 cm apart. To facilitate returning foragers' orientation and entry to their home colonies, each exit hole had a landing platform and the surrounding panel was painted with differing coloured patterns. To alleviate effects of adverse weather, we provided sugar water and pollen until day 14 after each colony's connection to the exterior or until 60 workers were present. Colonies still foraged externally during this period. The experiment ended when all adults within colonies had died naturally, which occurred in all cases by 21 August 2003.

Behavioural observations and sampling of bees

We inspected colonies daily and marked all newly eclosed workers on the thorax with colony-specific colour marks using water-based paint pens (mean $\pm$ s.d. = 82.4 ± 37.8 workers marked per colony). Counts were made of resident workers, colour-marked drifter workers and unmarked workers. All sexuals produced (140 queens and 7,916 males) were removed as newly eclosed adults and frozen. We conducted daily observation bouts (mean duration $\pm$ s.d. = 18.7 ± 7.8 min per day over 58.8 ± 19.1 days per colony) under red light to detect the competition point (outbreak of resident worker egg-laying and aggression)^{13,17}. To quantify their numbers, we colour-marked aggressive and egg-laying workers on the abdomen.

Genetic sampling and analyses

Before connecting nest boxes to the exterior, we took non-lethal genetic samples of resident workers by removing a tarsal tip²⁹. We supplemented this sample of queen's diploid offspring with colour-marked workers collected dead in their natal colonies and newly eclosed queens (Table 1). We genotyped colony queens and the contents of their sperm receptacles at the end of the experiment. We genotyped all samples (Table 1) at up to

six unlinked microsatellite loci (B10, B11, B96, B118, B121 and B124)^{10,13,30} using an Applied Biosystems 377 sequencer¹³. Loci had a mean of 9.5 alleles per locus (range 3–18) and a mean heterozygosity of 0.65. To prevent confusion with males from the external environment entering nest boxes, we only genotyped males that were unequivocally newly eclosed individuals ('callows'). To ensure accuracy of parentage attributions, all colony queens and (at diagnostic loci) putative worker-produced males were genotyped twice. We reconstructed each colony's mating type from genotypes of colony queens, their female offspring and their mates' sperm (Table 1). As expected^{10,11}, all female offspring had genotypes consistent with single queen mating. No inbreeding or diploid males were detected. Grouping the 32 colony queens into full sisterhoods (with COLONY²⁰) suggested that they originated from at least 25 different colonies (18 providing 1 queen and 7 providing 2 queens each), and hence, given the lack of population viscosity in *B. terrestris*²⁰, that colonies were, on average, unrelated.

Identification of worker-produced males

Given that only males produced by a queen's worker offspring can inherit an allele from her mate⁸, we selected within each colony 1–4 loci diagnostic of worker-produced males. We defined males as produced by resident workers if they possessed an allele of the queen's mate at one or more diagnostic loci (mean $\pm$ s.d. = 1.6 ± 0.8 loci with such alleles, $n = 28$ males). We defined males as produced by drifter workers if they possessed, at any locus, an allele absent from the queen or her mate (2.7 ± 1.3 loci with such alleles, $n = 53$ males). Males identified as produced by drifter workers were retyped at all loci to attribute them to their parents' colonies of origin. The number of males produced by resident workers was corrected to account for half such males lacking an allele of the queen's mate at diagnostic loci⁸. Within each phase of sampling (Table 1), we estimated overall numbers of males produced by queens, resident workers and drifter workers by multiplying their frequencies by the total number of males produced. To calculate per capita reproduction, we divided by the maximum number of resident or drifter workers present.

Received 11 May; accepted 17 June 2004; doi:10.1038/nature02769.

- Keller, L. (ed.) *Levels of Selection in Evolution* (Princeton University Press, Princeton, 1999).
- Queller, D. C. Relatedness and the fraternal major transitions. *Phil. Trans. R. Soc. Lond. B* **355**, 1647–1655 (2000).
- Sundström, L. & Boomsma, J. J. Conflicts and alliances in insect families. *Heredity* **86**, 515–521 (2001).
- Bourke, A. F. G. Worker reproduction in the higher eusocial Hymenoptera. *Q. Rev. Biol.* **63**, 291–311 (1988).
- Hamilton, W. D. The genetical evolution of social behaviour I, II. *J. Theor. Biol.* **7**, 1–52 (1964).
- Ratnieks, F. L. W. & Reeve, H. K. Conflict in single-queen Hymenopteran societies: the structure of conflict and processes that reduce conflict in advanced eusocial species. *J. Theor. Biol.* **158**, 33–65 (1992).
- Foster, K. R. & Ratnieks, F. L. W. Facultative worker policing in a wasp. *Nature* **407**, 692–693 (2000).
- Hammond, R. L., Bruford, M. W. & Bourke, A. F. G. Male parentage does not vary with colony kin structure in a multiple-queen ant. *J. Evol. Biol.* **16**, 446–455 (2003).
- Töth, E., Queller, D. C., Dollin, A. & Strassmann, J. E. Conflict over male parentage in stingless bees. *Insectes Soc.* **51**, 1–11 (2004).
- Estoup, A., Scholl, A., Pouvreau, A. & Solignac, M. Monoandry and polyandry in bumble bees (Hymenoptera: Bombinae) as evidenced by highly variable microsatellites. *Mol. Ecol.* **4**, 89–93 (1995).
- Schmid-Hempel, R. & Schmid-Hempel, P. Female mating frequencies in *Bombus* spp. from Central Europe. *Insectes Soc.* **47**, 36–41 (2000).
- Goulson, D. *Bumblebees: Their Behaviour and Ecology* (Oxford University Press, Oxford, 2003).
- Lopez-Vaamonde, C., Koning, J. W., Jordan, W. C. & Bourke, A. F. G. No evidence that reproductive bumblebee workers reduce the production of new queens. *Anim. Behav.* **66**, 577–584 (2003).
- Bulmer, M. G. Worker-queen conflict in annual social Hymenoptera. *J. Theor. Biol.* **93**, 239–251 (1981).
- Bourke, A. F. G. & Ratnieks, F. L. W. Kin-selected conflict in the bumble-bee *Bombus terrestris* (Hymenoptera: Apidae). *Proc. R. Soc. Lond. B* **268**, 347–355 (2001).
- Van Honk, C. G. J., Röseler, P.-F., Velthuis, H. H. W. & Hoogveen, J. C. Factors influencing the egg-laying of workers in a captive *Bombus terrestris* colony. *Behav. Ecol. Sociobiol.* **9**, 9–14 (1981).
- Duchateau, M. J. & Velthuis, H. H. W. Development and reproductive strategies in *Bombus terrestris* colonies. *Behaviour* **107**, 186–207 (1988).
- Duchateau, M. J. *Proc. XXth Int. Cong. Entomol* 404 (Florence, Italy, 1996).
- Moritz, R. F. A. & Neumann, P. Differences in nestmate recognition for drones and workers in the honeybee, *Apis mellifera* (L.). *Anim. Behav.* **67**, 681–688 (2004).
- Chapman, R. E., Wang, J. & Bourke, A. F. G. Genetic analysis of spatial foraging patterns and resource sharing in bumble bee pollinators. *Mol. Ecol.* **12**, 2801–2808 (2003).
- Bloch, G. Regulation of queen-worker conflict in bumble-bee (*Bombus terrestris*) colonies. *Proc. R. Soc. Lond. B* **266**, 2465–2469 (1999).
- Harder, L. D. Influences on the density and dispersion of bumble bee nests (Hymenoptera: Apidae). *Hol. Ecol.* **9**, 99–103 (1986).
- Paxton, R. J., Thorén, P. A., Estoup, A. & Tengö, J. Queen-worker conflict over male production and the sex ratio in a facultatively polyandrous bumblebee, *Bombus hypnorum*: the consequences of nest usurpation. *Mol. Ecol.* **10**, 2489–2498 (2001).
- Akre, R. D., Garnett, W. B., MacDonald, J. F., Greene, A. & Landolt, P. Behavior and colony development of *Vespula pensylvanica* and *V. atropilosa* (Hymenoptera: Vespidae). *J. Kans. Entomol. Soc.* **49**, 63–84 (1976).
- Martin, S. J., Beekman, M., Wossler, T. C. & Ratnieks, F. L. W. Parasitic Cape honeybee workers, *Apis mellifera capensis*, evade policing. *Nature* **415**, 163–165 (2002).
- Bourke, A. F. G. Sex ratios in bumble bees. *Phil. Trans. R. Soc. Lond. B* **352**, 1921–1933 (1997).
- Owen, R. E. & Plowright, R. C. Worker-queen conflict and male parentage in bumble bees. *Behav. Ecol. Sociobiol.* **11**, 91–99 (1982).
- Brown, M. J. F., Schmid-Hempel, R. & Schmid-Hempel, P. Queen-controlled sex ratios and worker reproduction in the bumble bee *Bombus hypnorum*, as revealed by microsatellites. *Mol. Ecol.* **12**, 1599–1605 (2003).
- Holehouse, K. A., Hammond, R. L. & Bourke, A. F. G. Non-lethal sampling of DNA from bumble bees for conservation genetics. *Insectes Soc.* **50**, 277–285 (2003).

- Estoup, A., Solignac, M., Cornuet, J.-M., Goudet, J. & Scholl, A. Genetic differentiation of continental and island populations of *Bombus terrestris* (Hymenoptera: Apidae) in Europe. *Mol. Ecol.* **5**, 19–31 (1996).

Acknowledgements We thank J. Lopez-Vaamonde and O. Rodriguez Ramos for help with the observations and processing of samples, J. J. M. Pereboom for advice, and T. Chapman and J. Field for comments on the manuscript. This work was funded by a NERC Research Grant to A.F.G.B. and W.C.J.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.L.-V. (carlos.lopez.vaamonde@ioz.ac.uk) or A.F.G.B. (andrew.bourke@ioz.ac.uk).

Neurons compute internal models of the physical laws of motion

Dora E. Angelaki, Aasef G. Shaikh, Andrea M. Green & J. David Dickman

Department of Neurobiology, Washington University School of Medicine, Saint Louis, Missouri 63110, USA

A critical step in self-motion perception and spatial awareness is the integration of motion cues from multiple sensory organs that individually do not provide an accurate representation of the physical world. One of the best-studied sensory ambiguities is found in visual processing, and arises because of the inherent uncertainty in detecting the motion direction of an untextured contour moving within a small aperture^{1–4}. A similar sensory ambiguity arises in identifying the actual motion associated with linear accelerations sensed by the otolith organs in the inner ear^{5,6}. These internal linear accelerometers respond identically during translational motion (for example, running forward) and gravitational accelerations experienced as we reorient the head relative to gravity (that is, head tilt). Using new stimulus combinations, we identify here cerebellar and brainstem motion-sensitive neurons that compute a solution to the inertial motion detection problem. We show that the firing rates of these populations of neurons reflect the computations necessary to construct an internal model representation of the physical equations of motion.

According to Einstein's equivalence principle⁷, linear accelerometers, including the otolith organs of the inner ear, detect net acceleration regardless of its source. For example, when we step to the side, primary otolith afferent neurons that encode the acceleration resulting from translational motion in the horizontal plane modulate their firing rates, as illustrated in Fig. 1a for a typical cell recorded in a macaque monkey. If instead of moving laterally we tilt our head to the opposite side, the afferent neurons change their firing rates in a similar way, this time to reflect the head realignment with respect to gravity (Fig. 1b). In fact, if the translational motion and tilt trajectories are appropriately matched in amplitude (as in Fig. 1; see also ref. 8), sensory otolith afferents will respond identically to the two stimuli^{5,6}.

When the displacement and tilt occur simultaneously, as illustrated in Fig. 1c, d, the firing rates of sensory afferent cells depend on the relative directions of the two types of movement. Whenever a head tilt to the right occurs simultaneously with a translational acceleration to the right, gravitational and inertial accelerations are oppositely directed and, if appropriately matched in amplitude, cancel each other out and neither component is transduced to the brain (Fig. 1c). Whenever a head tilt to the right occurs simultaneously with a translational acceleration to the left, the two

accelerations sum, resulting in a cell modulation that is double that of each movement alone (Fig. 1d).

Thus, although primary otolith afferents provide direction-selective movement information, the central nervous system is faced with the problem of identifying the type of motion that generated the sensory signal. In recent years, it has been proposed that the brain resolves this ambiguity by explicitly solving the equations of motion to construct central estimates of gravity and inertial (that is, translational) motion^{8–13}. Consistent with this hypothesis, velocity information arising from a different set of sensors, the semicircular canals, could be used to keep track of the rate of change of the gravity vector in the coordinate frame of the head (or equivalently the change in head orientation relative to gravity) according to the following equation¹⁴:

$$d\mathbf{g}/dt = -\boldsymbol{\omega} \times \mathbf{g} \quad (1)$$

where $\boldsymbol{\omega}$, $\mathbf{g}$ and $d\mathbf{g}/dt$ are vector representations of angular velocity, gravity and the rate of change of the gravitational acceleration, respectively, and ' $\times$ ' denotes the vector cross-product. The gravitational acceleration can be obtained by solving equation (1) (that is, $\mathbf{g} = -\int \boldsymbol{\omega} \times \mathbf{g}$). Because the net linear acceleration detected by the otolith sensors is equal to the vector difference of translational and gravitational components (that is, $\mathbf{a} = \mathbf{t} - \mathbf{g}$), a neural estimate of inertial motion acceleration, $\mathbf{t}$, can be computed as^{8–13}:

$$\mathbf{t} = \mathbf{a} - \int \boldsymbol{\omega} \times \mathbf{g} \quad (2)$$

For the specific stimuli of Fig. 1—that is, upright translation along the lateral (y axis) and rotation about the roll (x axis)—the task is simplified to computing the y -axis translational acceleration from the following scalar equation (Supplementary Fig. 1):

$$t_y = a_y + g_y = a_y + \int \omega_x g_z \quad (3)$$

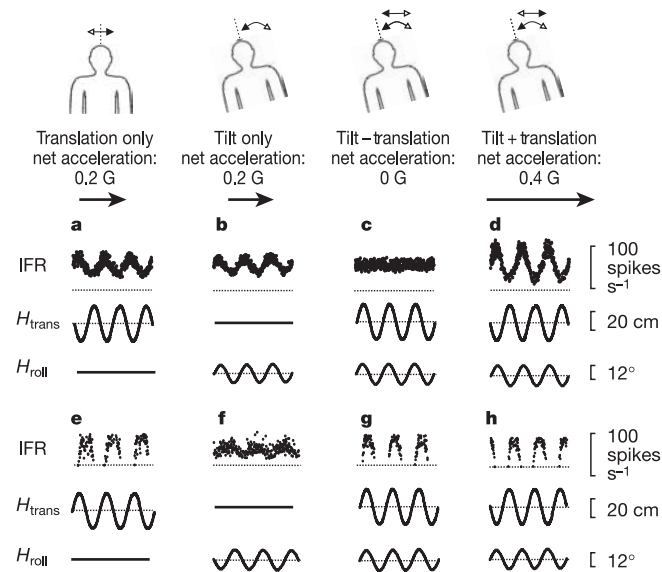


Figure 1 Neural responses that encode either the net gravito-inertial acceleration or the inertial component of linear acceleration. **a–d**, Instantaneous firing rates (IFR) of a primary otolith afferent (**a–d**) and a central VN cell (**e–h**) during: translation-only, tilt-only, tilt + translation and tilt + translation protocols (0.5 Hz). Straight arrows (schematic top) denote translational acceleration. Curved arrows represent deviations in tilt position. The stimulus (bottom) traces show translation position (H_{trans} , leftward positive) and tilt position (H_{roll} , right-ear-down positive). Data were recorded in a rhesus monkey during motion in darkness (see refs 6, 8 for details).

Equation (3) states that the neural estimate of the inertial acceleration along the y axis, t_y , can be approximated by a linear superposition of the net acceleration signal, a_y (sensed by the otoliths), and an internal estimate of the y -axis gravitational acceleration, $g_y = \int \omega_x g_z$. For small head movements from upright, $g_z \approx -1 \text{ G}$ ($G = 9.81 \text{ ms}^{-2}$) and $\int \omega_x g_z \approx -\int \omega_x$ such that g_y can be approximated as the integral of angular velocity sensed by the vertical semicircular canals^{12,13}.

Does the brain explicitly perform the computation implied by equation (3) to detect the accelerations associated with inertial motion? If so, neuron populations should exist whose firing rates are consistent with those predicted by the terms in this equation. We searched for a neural correlate of the proposed internal computations in the neural activities of two brain areas: the vestibular nuclei (VN) and the rostral fastigial nucleus of the cerebellum (FN). These areas not only receive sensory vestibular signals^{15–17} but also have projections suitable to distribute motion-related information for both perceptual and motor purposes^{8–12} through ascending projections to the thalamus/cortex and descending projections to the spinal cord^{18–22}. We used the stimuli in Fig. 1 to investigate FN and VN neuronal responses in alert monkeys, with the goal of identifying populations of cells that participate in the computations necessary to detect inertial motion and orientation in space.

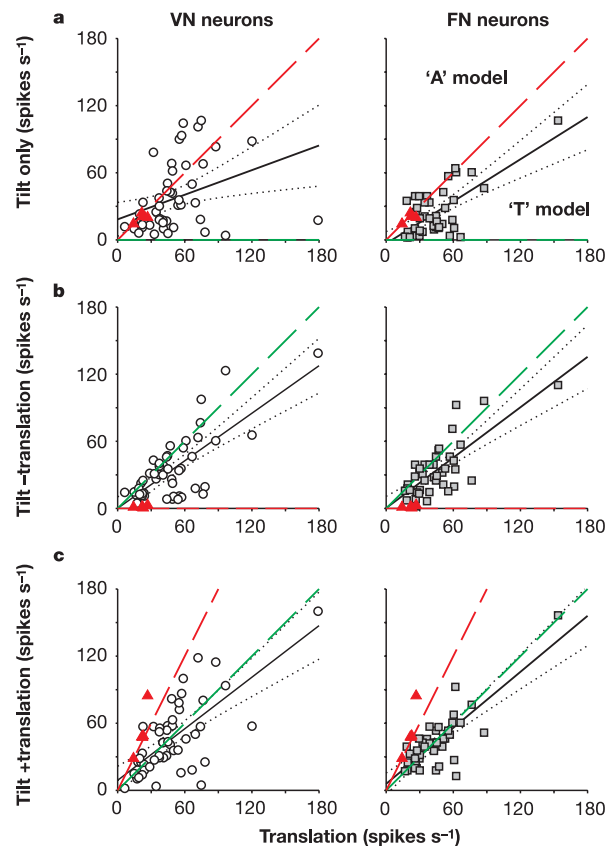


Figure 2 Summary of peak firing rates during **(a)** tilt-only and **(b, c)** the combined protocols as a function of the respective responses during translation-only motion. Each symbol corresponds to a cell (VN neurons, open circles; FN neurons, grey squares; otolith afferents⁶, red triangles). The predictions of two models have been illustrated with dashed lines: the 'T' model (green lines) corresponds to encoding the inertial (translational) motion acceleration and the 'A' model (red lines) illustrates the prediction for afferent-like neurons encoding the net gravito-inertial acceleration. Solid lines illustrate linear regressions (dotted lines are 95% confidence intervals).

Responses from a VN cell that encoded inertial motion are illustrated in Fig. 1e–h. Like the sensory afferent cell of Fig. 1a–d, this neuron was directionally selective with a preference for lateral movements (Fig. 1e). However, when the same acceleration was instead generated by a head tilt, this central neuron's peak modulation was only 10% of its response during translation (Fig. 1f). Combined motion protocols (Fig. 1g, h) resulted in cell activation that was similar to that during pure translation (Fig. 1e), consistent with the notion that the brain has computed an internal representation of translational motion.

Other VN and FN neurons exhibited intermediate properties, as summarized in Fig. 2. During tilt, translation-selective neurons should not modulate and peak firing rates should be zero (Fig. 2a; the green dashed lines falling along the abscissa represent the prediction of a translation, here called the 'T' model). At the other extreme, sensory-like neurons should modulate similarly during tilt and translation, with peak firing rates falling along the unity-slope line (Fig. 2a; the red dashed lines represent the prediction of a net acceleration, the 'A' model). Most (66 out of 93) VN and FN cells modulated less during tilts, compared with translation (Fig. 2a; open and grey symbols falling below the unity-slope red lines). However, in only 5 out of 93 neurons was tilt modulation less than 10% of the peak response amplitude during translation,

indicating that few cells responded as expected for purely translation-selective neurons (Fig. 2a; green lines).

For the combined movement trajectories, the responses of translation-selective neurons should be similar to those during translation-only motion, with peak firing rates that fall along the unity-slope green line (Fig. 2b, c). Although some cells had peak firing rates similar to those measured during translation, responses recorded during the combined protocols were generally different than those recorded during translation-only motion ($F_{(2,182)} = 15.4$, $P < 0.01$). However, most cells did not encode the net acceleration either (Fig. 2b, zero slope; Fig. 2c, slope of 2, red lines). Translational motion is oppositely directed during the combined movement protocols (that is, tilt – translation and tilt + translation motions). Thus, the observation that the peak response modulation was typically also reversed in the two cases provided further evidence that central neurons reflect processed, rather than sensory-like, motion information (Supplementary Fig. 2).

We used linear regression analyses to quantify whether the afferent-like, net acceleration ('A' model) or the translation ('T' model) was most appropriate to describe mean firing rate modulation during all four stimulus combinations. To make this assessment under conditions in which a given neuron could be significantly correlated with both models we used a partial correlation analysis³. Figure 3a shows a scatter plot of the z-score equivalents of the partial correlation coefficients calculated for each cell. The 'T' model provided significantly ($P < 0.01$) better fits than the 'A' model in two-thirds (28 out of 42, 67%) of FN cells but only 37% (19 out of 51) of VN neurons (data in the upper-left quadrant of Fig. 3a). The reverse was true in 28% of FN cells and 41% of VN neurons (lower-right quadrant). The fact that cerebellar FN neuron responses were closer to coding translational motion, when compared with brainstem VN cells, is further emphasized by significant differences in the distribution of r^2 values for each population ($X^2 = 78$, d.f. = 9, $P < 0.01$). However, because only 4 VN and 13 FN neurons had $r^2 > 0.9$ (Fig. 3b), the majority of brainstem and cerebellar motion-sensitive neurons clearly encode neither the sensory net acceleration signal nor an internal estimate of inertial motion. What then do the firing rates of these cells encode?

We next tested the hypothesis that central motion-sensitive neuron properties represent a distributed coding of the mathematical terms on the right-hand side of equation (3). The latter, referred to as the 'sensory convergence' model, embeds terms related to net acceleration (sensed by the otoliths) and an internal estimate of gravitational acceleration (derived from semicircular canal signals); that is, the 'C' model ($k_1 a_y + k_2 \int \omega_x g_z$). However, in contrast to the 'T' model, in which $k_1 = k_2$ (equation (3)), the 'C' model allows for unequal coefficients, in which k_1 and k_2 are complex numbers that embed information about both response amplitude and phase (see Methods). Figure 4a, b compares the distributions of correlation coefficients for the 'C' model with those of the 'A' model. The 'C' model accounted for the patterns of activation and provided significantly better fits ($P < 0.01$, sequential F -test) in 84 out of 93 VN and FN neurons.

If individual neurons coded explicitly for translation, equation (3) implies that coefficients k_1 and k_2 of the 'C' model should have identical amplitude and phase values. However, for most cells $k_1 \neq k_2$, providing an explanation as to why individual cells did not all comply with the 'T' model predictions. Nonetheless we found that the amplitudes of the respective coefficients, k_1 and k_2 , were linearly correlated for both VN and FN neurons, with population slopes that were not statistically different from unity ($P > 0.01$; Fig. 4c, top). Thus, k_1 and k_2 were equally represented in the population activities, suggesting that translational motion information is centrally encoded in a distributed fashion.

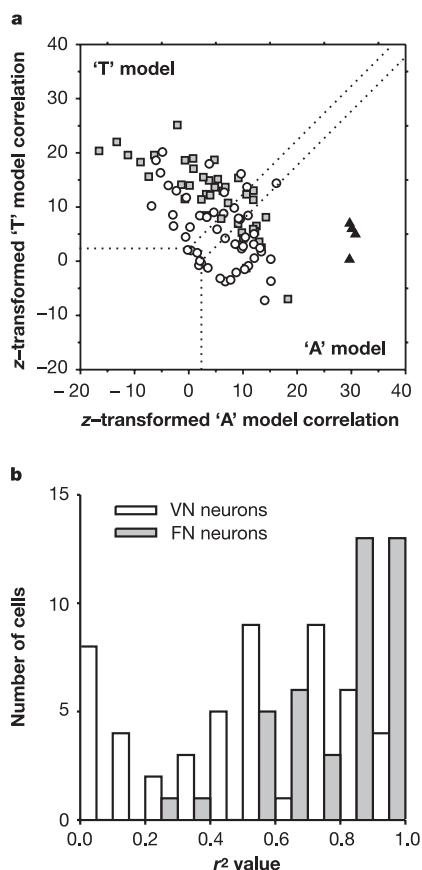


Figure 3 Cerebellar FN neurons more closely encode the inertial component of acceleration than does the brainstem VN cell population. **a**, Scatter plot of z-scores corresponding to the partial correlation coefficients for the 'A' (abscissa) and 'T' model (ordinate) fits to each neuron. Dotted lines define regions where: (1) the 'A' model provided significantly better fits than the 'T' model ($z > 2.33$, $P < 0.01$; area below and to the right of dotted lines); (2) the 'T' model provided better fits than the 'A' model (area above and to the left of dotted lines); (3) the cells were not selective for either model (between the dotted lines). **b**, Values of r^2 for 'T' model fits. VN neurons, open circles/bars; FN, grey squares/bars; otolith afferents, black triangles.

The phase of the two coefficients further supports this conclusion. The regression analysis imposed no constraints on the relative phase distributions for k_1 and k_2 . Nonetheless, we found that the phases associated with these coefficients co-varied, with slopes that were not statistically different from unity ($P > 0.01$; Fig. 4c, bottom). Individual neurons thus tended to encode net and gravitational acceleration components with similar phases. Importantly, although otolith afferents encode net linear acceleration (0° phase for k_1 coefficient), the internal estimate of gravity is probably derived from the semicircular canals^{8,12,23}, which encode angular velocity²⁴ ($\pm 90^\circ$ for k_2 coefficient). However, neural responses did not simply reflect these distinct afferent dynamics; rather, despite a large distribution of temporal dynamics across the cell population^{6,25}, the phases of the k_1 and k_2 coefficients were appropriately matched. Thus, dynamically processed sensory information from the different sensors has converged in these neurons as required to construct a distributed, internal model representation of inertial motion (equation (3)). Subsequent neural processing could use a simple weighted population average of responses to extract inertial motion information at the level of individual neurons for both perception and eye movements^{8,9,12}.

These results illustrate a direct correlation between cell firing rates and the equations of motion, as applied to movement in a gravitational environment and the physics of the external world. A neural basis for an internal model representation of the relationship between the physical environment and either the sensory detectors or the motor apparatus has only recently begun to be explored^{26–28}. Here we have shown evidence that, in support of theoretical predictions^{8–13}, subcortical neural populations might provide a distributed solution to the inertial motion detection problem. □

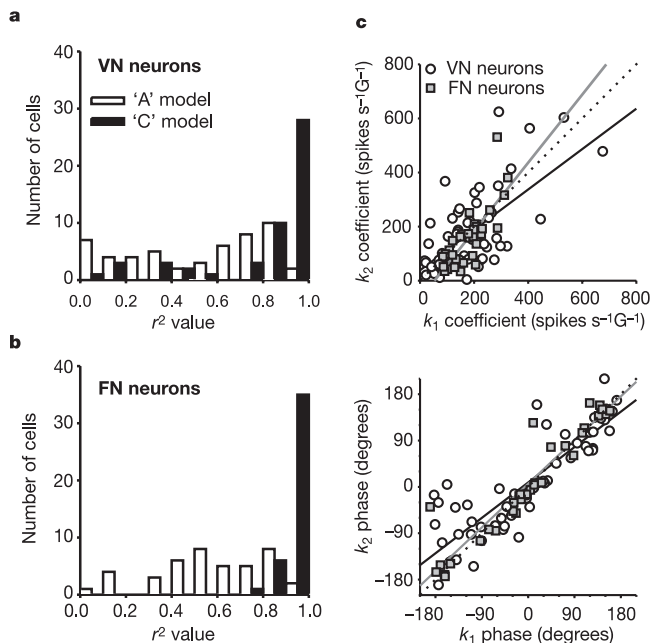


Figure 4 An internal model of equation (3) accounts best for the modulation in neural firing rates. **a, b**, Bar histograms for the r^2 values corresponding to fits of the 'A' and 'C' models (open and filled, respectively). **c**, Relationship between the k_1 and k_2 coefficients (decomposed into amplitude and phase) for each cell (VN neurons, open circles; FN neurons, filled squares). Black (VN cells) and grey (FN cells) solid lines illustrate linear regressions for the amplitude (VN: $r^2 = 0.42$, $P < 0.01$ (slope = 0.75); FN: $r^2 = 0.60$, $P < 0.01$ (slope = 1.26)) and phase (VN: $r^2 = 0.76$, $P < 0.01$ (slope = 0.76); FN: $r^2 = 0.88$, $P < 0.01$ (slope = 0.94)). Dotted lines are unity-slope lines.

Methods

Neural activities in the VN and the FN were recorded in macaque monkeys (one *Macaca fascicularis* and two *Macaca mulata*) using procedures similar to those in previous studies^{6,25,29}. All surgical procedures were performed under sterile conditions in accordance with institutional and the NIH guidelines. To minimize extra-vestibular sensory influences, animals were passively moved in complete darkness with the head fixed relative to the body. Only cells responsive to translational motion and whose firing rates were not correlated with either slow or fast eye movements have been characterized with the experimental protocol illustrated in Fig. 1 (top). Briefly, using a motion delivery system consisting of a rotator on top of a translator, four stimulus combinations of 0.5 Hz translational and rotational motion stimuli were used⁸: Translation-only, tilt-only, or combined translation and tilt movements (tilt + translation motion and tilt - translation motion). The tilt stimulus consisted of a $\pm 11.3^\circ$ ($35.5^\circ \text{ s}^{-1}$) rotation from upright. As a result of such a change in head and body orientation relative to gravity, a horizontal sinusoidal linear acceleration (gravity) component with a peak magnitude of $\pm 0.2 \text{ G}$ stimulated the otoliths along the interaural head axis. The amplitude of the translation stimulus was then adjusted ($\pm 20 \text{ cm}$) such that peak linear acceleration matched that induced by the head reorientation relative to gravity during the tilting movement. As a result, during combined rotational and translational stimulation the net acceleration activating the otolith receptors either doubled (tilt + translation) or was nearly zero (tilt - translation), even though the actual translation of the animal remained the same (Fig. 1). These motion protocols were delivered along the preferred translational motion direction for each cell.

The instantaneous firing rate was computed off-line as the inverse of interspike interval, and data from multiple motion repetitions were folded into a single cycle that was fitted by a sine function^{6,25,29}. To determine the physical parameters encoded by each cell, linear regression analyses were used²⁹ to simultaneously fit the cumulative cycles of cell modulation during each of the translation, tilt and combined stimuli (Fig. 1) using the following models: (1) two first-order models (the 'T' model and the 'A' model), which assume that neural firing rates could be modelled as coding either the translational acceleration component or the net acceleration, respectively; and (2) a sensory convergence model consistent with a distributed representation of equation (3) (that is, the 'C' model = $k_1 a_y + k_2 \int \omega_x g_z$), which assumes that neural firing rates could be modelled as a linear combination of net acceleration, (a_y , sensed by the otoliths) and an internal estimate of gravity ($g_y = \int \omega_x g_z$, derived from semicircular canal signals).

For small amplitude tilt movements, $\int \omega_x g_z = 0$ during translation-only and $\int \omega_x g_z = -\theta_0(1 - \theta_0^2/8) \sin(2\pi ft)$ for tilt-only and combination profiles, in which $\theta(t) = \theta_0 \sin(2\pi ft)$ describes the tilt position (in units of radians) and $f = 0.5 \text{ Hz}$. For the relatively small tilt amplitudes used here, $\int \omega_x g_z \approx -\theta_0 \sin(2\pi ft)$, approximating tilt position¹². Note that k_1 and k_2 are complex coefficients associated with both amplitude and phase values (computed according to the procedure described in ref. 29).

How well each of the 'T' and 'A' models (same number of parameters) fitted the data was evaluated using a partial correlation analysis. Partial correlation coefficients³ were subsequently converted to z-scores using Fisher's r -to- z transform to facilitate the interpretation of statistical significance independently of the number of data points. For the 'C' model, the goodness of fit was evaluated using the correlation coefficients and a sequential F -test³⁰ to compare the errors associated with each of the 'C' and 'A' models while allowing for differences in their degrees of freedom. The specificity of these models to account for the firing rate modulation of the cells was also verified using a permutation test, in which the response profiles of each cell for the translation-only, tilt-only, tilt - translation and tilt + translation conditions were randomly reassigned many times and the r^2 values for the model fit to the permuted responses compared with those of the original data set. Statistical analyses of neural responses used analyses of variance with repeated measures and histogram comparisons were based on chi-square tests.

Received 12 December 2003; accepted 14 June 2004; doi:10.1038/nature02754.

- Shimojo, S., Silverman, G. H. & Nakayama, K. Occlusion and the solution to the aperture problem for motion. *Vision Res.* **29**, 619–626 (1989).
- Adelson, E. H. & Movshon, J. A. Phenomenal coherence of moving visual patterns. *Nature* **300**, 523–525 (1982).
- Movshon, J. A., Adelson, E. H., Gizzi, M. S. & Newsome, W. T. in *Pattern Recognition Mechanisms: Pontificiae Academiae Scientiarum Scripta Varia* (eds Chagas, C., Gattass, R. & Gross, C.) Vol. 54, 117–151 (Vatican, Rome, 1985).
- Pack, C. C. & Born, R. T. Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature* **409**, 1040–1042 (2001).
- Fernández, C. & Goldberg, J. M. Physiology of peripheral neurons innervating otolith organs of the squirrel monkey. I. Response to static tilts and to long-duration centrifugal force. *J. Neurophysiol.* **39**, 970–984 (1976).
- Angelaki, D. E. & Dickman, J. D. Spatiotemporal processing of linear acceleration: primary afferent and central vestibular neuron responses. *J. Neurophysiol.* **84**, 2113–2132 (2000).
- Einstein, A. Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungen. *Jber. Radioakt.* **4**, 411–462 (1908).
- Angelaki, D. E., McHenry, M. Q., Dickman, J. D., Newlands, S. D. & Hess, B. J. M. Computation of inertial motion: neural strategies to resolve ambiguous otolith information. *J. Neurosci.* **19**, 316–327 (1999).
- Merfeld, D. M., Zupan, L. H. & Peterka, R. J. Humans use internal models to estimate gravity and linear acceleration. *Nature* **398**, 615–618 (1999).
- Merfeld, D. M. & Zupan, L. H. Neural processing of gravito-inertial cues in humans. III. Modeling tilt and translation responses. *J. Neurophysiol.* **87**, 819–833 (2002).
- Zupan, L. H., Peterka, R. J. & Merfeld, D. M. Neural processing of gravito-inertial cues in humans. II. Influence of the semicircular canals following post-rotatory tilt. *J. Neurophysiol.* **84**, 2001–2015 (2000).
- Green, A. M. & Angelaki, D. E. Resolution of sensory ambiguities for gaze stabilization requires a second neural integrator. *J. Neurosci.* **23**, 9265–9275 (2003).

13. Green, A. M. & Angelaki, D. E. An integrative neural network for detecting inertial motion and head orientation. *J. Neurophysiol.* Advance online publication, 31 March 2004 (doi: 10.1152/jn.01234.2003).
14. Goldstein, H. *Classical Mechanics* (Addison-Wesley, Reading, Massachusetts, 1980).
15. Carleton, S. C. & Carpenter, M. B. Distribution of primary vestibular fibers in the brainstem and cerebellum of the monkey. *Brain Res.* **294**, 281–298 (1984).
16. Newlands, S. D. *et al.* Central projections of the saccular and utricular nerves in macaques. *J. Comp. Neurol.* **466**, 31–47 (2003).
17. Kotchabhakdi, N. & Walberg, F. Primary vestibular afferent projections to the cerebellum as demonstrated by retrograde axonal transport of horseradish peroxidase. *Brain Res.* **142**, 142–146 (1978).
18. Angaut, P. & Bowsler, D. Ascending projections of the medial cerebellar (fastigial) nucleus: an experimental study in the cat. *Brain Res.* **24**, 49–68 (1970).
19. Asamura, C., Thach, W. T. & Jones, E. G. Distribution of cerebellar terminations and their relation to other afferent terminations in the ventral lateral thalamic region of the monkey. *Brain Res. Rev.* **5**, 237–265 (1983).
20. Batton, R. R., Jayaraman, A., Ruggiero, D. & Carpenter, M. B. Fastigial efferent projections in the monkey: an autoradiographic study. *J. Comp. Neurol.* **174**, 281–305 (1977).
21. Kalil, K. Projections of the cerebellar and dorsal column nuclei upon the thalamus of the rhesus monkey. *J. Comp. Neurol.* **195**, 25–50 (1981).
22. Lang, W., Buttner-Ennever, J. A. & Buttner, U. Vestibular projections to the monkey thalamus: An autoradiographic study. *Brain Res.* **177**, 3–17 (1979).
23. Newlands, S. D., Shaikh, A. G., Green, A. M., Dickman, J. D., Angelaki, D. E. Semicircular canal inactivation abolishes tilt-translation discrimination property of fastigial and vestibular nuclei neurons. (Abstract viewer/itinerary planner, Society for Neuroscience, Washington DC, 2004)
24. Goldberg, J. M. & Fernandez, C. Physiology of peripheral neurons innervating semicircular canals of the squirrel monkey. II. Response to sinusoidal stimulation and dynamics of peripheral vestibular system. *J. Neurophysiol.* **34**, 661–675 (1971).
25. Dickman, J. D. & Angelaki, D. E. Vestibular convergence patterns in vestibular nuclei neurons of alert primates. *J. Neurophysiol.* **88**, 3518–3533 (2002).
26. Gribble, P. L. & Scott, S. H. Overlap of internal models in motor cortex for mechanical loads during reaching. *Nature* **417**, 938–941 (2002).
27. Kawato, M. Internal models for control and trajectory planning. *Curr. Opin. Neurobiol.* **6**, 718–727 (1999).
28. Padoa-Schioppa, C., Li, C.-S. R. & Bizzi, E. Neuronal correlates of kinematics-to-dynamics transformation in the supplementary motor area. *Neuron* **36**, 751–765 (2002).
29. Angelaki, D. E. & Dickman, J. D. Premotor neurons encode torsional eye velocity during smooth-pursuit eye movements. *J. Neurosci.* **23**, 2971–2979 (2003).
30. Draper, N. R. & Smith, H. *Applied Regression Analysis* (J. Wiley, Sons, Inc., New York, 1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The work was supported by grants from NIH and NASA. We thank E. Klier, C. Fetsch, G. DeAngelis and L. Snyder for aid in preparation of the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.A. (angelaki@pcg.wustl.edu).

Median bundle neurons coordinate behaviours during *Drosophila* male courtship

Devanand S. Manoli^{1,2} & Bruce S. Baker²

¹Neurosciences Program, ²Department of Biological Sciences, Stanford University, Stanford, California 94305, USA

Throughout the animal kingdom the innate nature of basic behaviour routines suggests that the underlying neuronal substrates necessary for their execution are genetically determined and developmentally programmed^{1–2}. Complex innate behaviours require proper timing and ordering of individual component behaviours. In *Drosophila melanogaster*, analyses of combinations of mutations of the *fruitless* (*fru*) gene have shown that male-specific isoforms (*Fru^M*) of the *Fru* transcription factor are necessary for proper execution of all steps of the innate courtship ritual^{3–9}. Here, we eliminate *Fru^M* expression in one group of about 60 neurons in the *Drosophila* central nervous

system and observe severely contracted courtship behaviour, including rapid courtship initiation, absence of orienting and tapping, and the simultaneous occurrence of wing vibration, licking and attempted copulation. Our results identify a small group of median bundle neurons, that in wild-type *Drosophila* appropriately trigger the sequential execution of the component behaviours that constitute the *Drosophila* courtship ritual.

Genetic, developmental and behavioural studies of the sex determination gene *fru* (Fig. 1) show that *Fru^M* proteins are only produced in a small, limited and dispersed subset of central nervous system (CNS) cells^{6,10,11}. These studies also suggest that *Fru^M* is responsible for building the potential for male courtship behaviour into the CNS during development, and that the cells that express *Fru^M* comprise the circuitry that controls and coordinates male courtship behaviours^{1,6,9}. Here we test the proposition that subsets of *Fru^M* cells have specific roles in courtship by examining the behavioural phenotypes produced when *Fru^M* expression is eliminated in a specific subset of CNS neurons.

We identified a *GAL4* enhancer trap line, *P52a-GAL4* (ref. 12), whose expression overlaps a bilateral cluster of ~60 *Fru^M*-expressing neurons in the suboesophageal ganglion (Fig. 2a, b; white). These neurons comprise part of the tritocerebral component of the median bundle. *P52a-GAL4* expression is not detected in any other *Fru^M*-expressing neurons, although it is expressed in other regions of the CNS (Fig. 2a, b; green). *P52a-GAL4*-directed expression of an RNA-mediated interference transgene (*UAS-fru^{MIR}*) targeting the male-specific amino terminus of *Fru^M* isoforms eliminates *Fru^M* expression in the ~60 *P52a*-labelled *Fru^M* neurons (Fig. 2c), and is not expected to affect the hundreds of other cells in which *P52a-GAL4* is expressed.

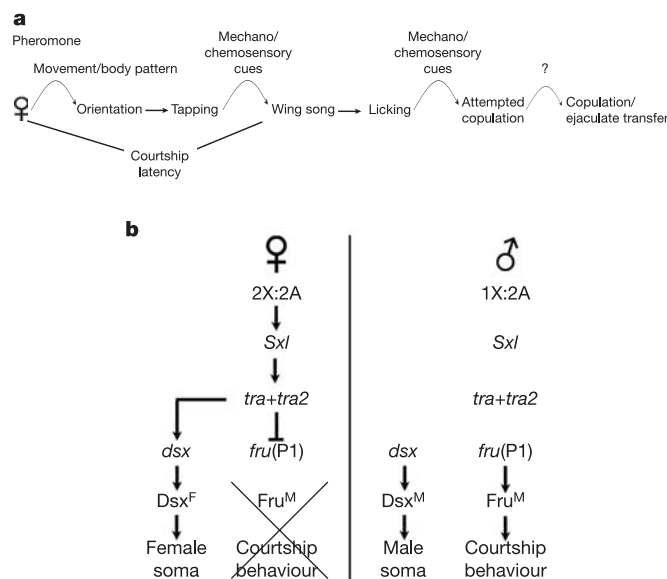


Figure 1 The *Drosophila* courtship ritual and its regulation by the *fru* branch of the *Drosophila* sex-determination hierarchy. **a**, Visual and olfactory cues used to detect and identify an appropriate female elicit orienting behaviour as the male faces and follows the female at a constant distance^{3–5}. The male then orients to the female's lateral aspect and 'taps' her abdomen, perhaps facilitating the proper recognition of females and progression through the courtship ritual. The male then extends his outer wing and vibrates it to generate a species-specific courtship song, reorients towards the female's posterior, and extends his proboscis to 'lick' her genital region. Following successful licking, the male will then curl his abdomen and attempt copulation. Copulation typically lasts for about 20 min and results in the transfer of sperm and seminal fluids. **b**, The regulation of male courtship through *Fru^M* by the *Drosophila* sex-determination hierarchy. In males, the lack of *Sxl* and thus *Tra* activity leads to the default splicing of *fru* P1-derived transcripts, yielding *Fru^M*.

Males in which Fru^M expression had been eliminated in median bundle neurons by the *P52a-GAL4*-directed expression of *UAS-fru^{MIR}* (*P52a/fru^{MIR}*) were used in standard courtship assays (see Methods) to assess the Fru^M -dependent roles of these neurons in courtship. In *P52a/fru^{MIR}* males, courtship latency—the period from the initial presentation of a virgin female to the initiation of courtship behaviour, defined here as wing extension; Fig. 1a—decreased (8 ± 1 s ($\pm$ s.e.m.) for *P52a/fru^{MIR}* males, compared with 94 ± 8 s for control males; Fig. 3a and Table 1). However, *P52a/fru^{MIR}* males can still distinguish females from males, because they do not sustain courtship towards each other or towards control males (data not shown), unlike previously described mutants that exhibited a rapid initiation of courtship towards both virgin females and mature males¹³. We did several controls to ensure that the rapid initiation of courtship seen in *P52a/fru^{MIR}* males is the consequence of blocking Fru^M expression in these 60 median bundle neurons. All of the individual transgenes used in these studies were backcrossed into a common genetic background before use. For each of these transgenes the courtship behaviours of males carrying that transgene alone did not differ from our controls (Fig. 3a). Additionally, the *P52a-GAL4*-directed expression of a *UAS-traF* transgene (Fig. 1b) also eliminates Fru^M expression in these 60 neurons (data not shown) and reduces courtship latency (10 ± 2 s versus 94 ± 8 s) (Fig. 3a and Table 1). On the basis of these and other controls (see Methods), we conclude that it is the elimination of Fru^M protein expression in the ~ 60 median bundle neurons, through the *P52a*-driven expression of *UAS-fru^{MIR}*, that is responsible for the decreased courtship latency.

To address whether rapid courtship by *P52a/fru^{MIR}* males was a reflection of general heightened activity, we performed short-, intermediate and long-term locomotor assays on both control and *P52a/fru^{MIR}* males¹⁴ (Table 1 and Fig. 3b). There were no significant differences in their activity (see Methods), suggesting that the behavioural differences observed in *P52a/fru^{MIR}* males are specific to courtship.

The longer courtship latency seen in wild-type relative to *P52a/fru^{MIR}* males suggests that initiation of courtship by wild-type males requires sensory information beyond that sufficient to elicit courtship by *P52a/fru^{MIR}* males. Whether this represents a need for qualitatively different information in the wild type (that is, specific sensory information not required to elicit courtship by *P52a/fru^{MIR}* males), or quantitative differences in the information obtained (that is, stimuli integrated over the longer latency period in the wild type), remains undetermined. These data suggest that in the wild type, Fru^M -dependent median bundle function inhibits the rapid

onset of courtship until sufficient stimuli overcome this inhibition (Fig. 4a, Supplementary Fig. 1), and that initiation requires both proper recognition of females—the cues sufficient to elicit courtship in *P52a/fru^{MIR}* males—as well as potentially a second stimulus, which acts to relieve the inhibition of courtship initiation (Fig. 4b).

P52a/fru^{MIR} males also display a marked compression of the remainder of the courtship ritual, quantified as the period from the initiation of courtship to attempted copulation. In the wild type, courtship typically involves multiple tapping interactions before progression to wing song and attempted licking, and often numerous licking attempts before attempted copulation. In control males, such courtship bouts extend on average for about 2 min (111 ± 20 s). In marked contrast, *P52a/fru^{MIR}* males attempt copulation within 4 s after courtship initiation (Table 1).

Analysis of videotaped behaviour of *P52a/fru^{MIR}* males revealed that the initial stages of courtship (orienting and tapping) were absent. These males rapidly progressed to the later stages of licking and attempted copulation, with wing vibration, proboscis extension and attempted copulation often occurring simultaneously (Fig. 3d, Table 1 and see below). Furthermore, proboscis extension was never directed towards the female's genitalia, but rather towards her dorsal posterior while copulation was being attempted, and never resulted in a successful contact of the proboscis with genital structures. Thus in *P52a/fru^{MIR}* males, the requirements for direct contact with females (tapping and licking) are bypassed, and the recognition of a female seems to immediately elicit attempted copulation.

Previous studies concluded that tapping is necessary for proper species and sexual discrimination^{3,4,15}. However, our findings with respect to *P52a/fru^{MIR}* males indicate that appropriate recognition of females can occur in its absence, implying that initial visual and/or olfactory cues are sufficient for the proper recognition of females. Moreover, these cues are able to elicit all components of the courtship ritual when the requirement for positively-acting cues received by means of tapping and licking are bypassed. The information obtained by tapping, although still potentially providing sex- and species-specific cues, may function as a necessary step in wild-type situations by facilitating progression from orienting and tracking to wing song and subsequent courtship behaviours.

The complex courtship phenotype of *P52a/fru^{MIR}* males—the rapidity of courtship initiation, absence of orienting and tapping, and simultaneous occurrence of later steps—suggests that Fru^M -expressing median bundle neurons function to modulate progression through courtship. There are at least two possible models for how this may occur. First, Manning has suggested that the

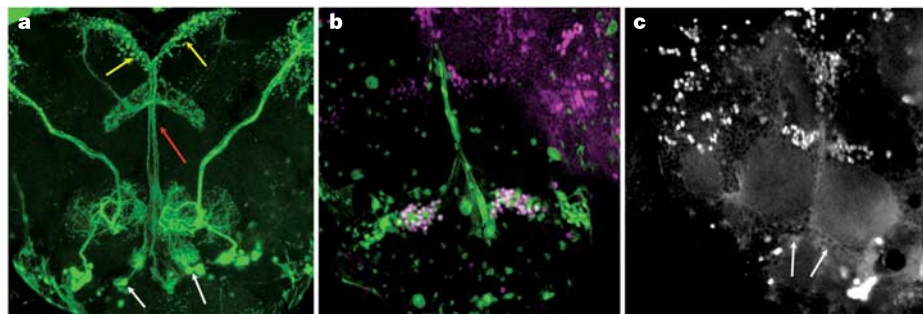


Figure 2 *P52a-GAL4* expression in Fru^M neurons of the median bundle. *P52a-GAL4* directs inhibition of Fru^M expression by a *UAS-fru^{MIR}* transgene. **a**, *P52a-GAL4* expression in the *Drosophila* brain is shown by expression of UAS-driven cytoplasmic β -Gal. Arrows indicate cell bodies of median bundle neurons in the suboesophageal ganglion (white), their projections dorsally in the median bundle (red) and their ramification in the dorsal protocerebrum (yellow). **b**, *P52a-GAL4* directs GFPnls (green)

expression in Fru^M -expressing (magenta; overlap in white) suboesophageal ganglion cells. View of the anterior surface. **c**, *P52a-GAL4*-directed expression of a *UAS-fru^{MIR}* transgene successfully inhibits Fru^M expression in suboesophageal ganglion neurons. Arrows indicate the region of the suboesophageal ganglion where Fru^M expression is absent (compare with **b**)

integration of excitatory and inhibitory cues over time contribute to increasing excitation in males⁵, and that progressively increasing levels of excitation then drives the progression through the courtship sequence, for example, by sequential components requiring progressively higher levels of excitation for their elicitation. Thus in wild-type males, Fru^M-dependent median bundle neuron function would be required for the system of progressively higher thresholds (Supplementary Fig. 1a). In *P52a/fru^MIR* males, the thresholds for the various steps in courtship would be substantially lower and equivalent, such that minimal stimulation could simultaneously elicit wing song, licking and attempted copulation (Supplementary Fig. 1b). An alternative model is that in the wild type, Fru^M-dependent median bundle function is necessary for tonic inhibitions that block progression from one step to the next in courtship. In the wild type, each of these tonic inhibitions can be relieved by the male receiving appropriate sensory cues (Fig. 4a). In *P52a/fru^MIR* males, such tonic inhibitions would be absent, and thus the later steps in courtship would be simultaneously triggered by recognition of a female (Fig. 4b).

Another behavioural anomaly in *P52a/fru^MIR* males occurs when multiple males are presented with a single female. In control

situations, if two or more males are present with a single female, initiation of courtship by one male delays the initiation of courtship by the second male (Table 1). When multiple *P52a/fru^MIR* males are with a single female, all males within the chamber immediately court and attempt to copulate with the female. The courtship by multiple *P52a/fru^MIR* males is distinct from the lines of males that court a single female at the head of the chain, as seen with certain *fru* alleles¹⁴, because all *P52a/fru^MIR* males focus on the posterior of the single female, consistent with defects in behaviour downstream of proper mate recognition (Supplementary Video 1). These data suggest that in wild-type situations, cues from one male or a male–female pair act through the median bundle to inhibit the initiation of courtship behaviour by a second male (Fig. 4c).

Given the advantage that vigorous, rapid-onset courtship might confer, and the related question of why such behaviour has not been selected for as the species norm, we examined copulation by and fertility of *P52a/fru^MIR* males. Whereas *P52a/fru^MIR* males attempt copulation more quickly, successful copulation occurred more rapidly in control males, with 8 out of 20 successful attempts within the first 5 min, compared with 0 out of 20 in *P52a/fru^MIR* males. Moreover, although copulations by *P52a/fru^MIR* males were of

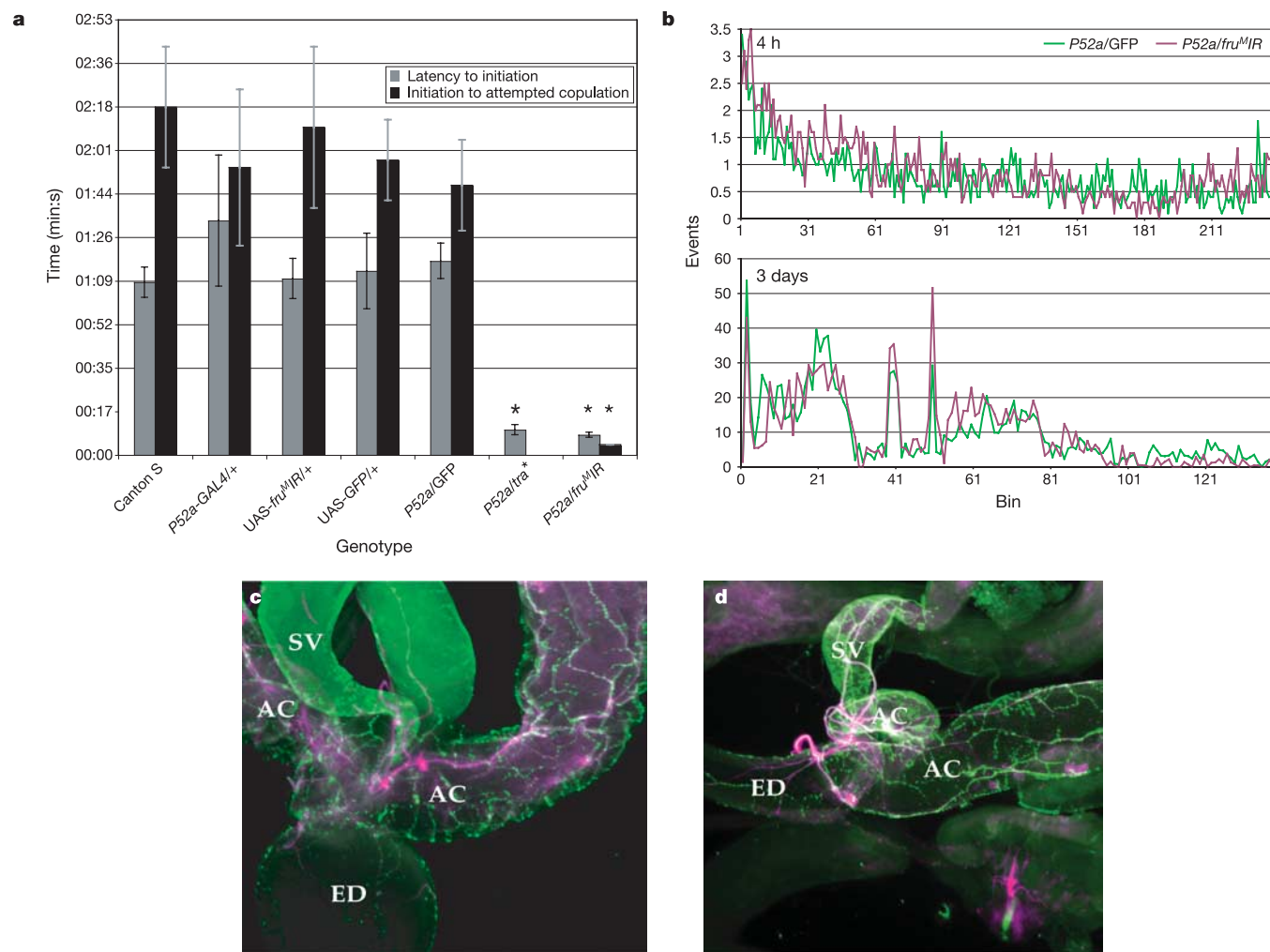


Figure 3 Inhibition of Fru^M expression in median bundle neurons leads to aberrant courtship behaviour. **a**, Average courtship latencies for control and *P52a*-manipulated males in which Fru^M expression has been reduced using directed expression of a *UAS-tra.F* (female) or *UAS-fru^MIR* transgene ($n = 15$ for Canton S, *P52a/+*, *UAS-fru^MIR/+*, *UAS-GFP* and *P52a/tra.F*, and 20 for *P52a/GFP* and *P52a/fru^MIR*). *P52a*-manipulated males show decreases, relative to control males, in courtship latency when presented with virgin females. An asterisk indicates $P < 0.01$ for values compared

with *P52a/GFP* controls. Error bars indicate s.e.m. values. **b**, *P52a/fru^MIR* (purple) and control (green) males show no differences in intermediate (top panel, $P = 0.79$) or long-term (bottom panel, $P = 0.65$) locomotor activity. **c**, **d**, Although sterile, *P52a/fru^MIR* males (**d**), compared with control *P52a/LacZ* (**c**), show no defects in either the innervation of internal genitalia (mAb 22C10 staining, magenta), or the serotonergic differentiation of abdominal ganglion neurons that innervate portions of the internal genitalia (anti-5HT staining, green). ED, ejaculatory duct; SV, seminal vesicle; AC, accessory gland.

normal duration (17.4 min versus Canton S 17.5), these matings were all sterile, despite the presence of motile sperm in their testes (Table 1). Because sperm and seminal fluid transfer are dependent on the functioning of eight Fru^M -expressing serotonergic neurons that provide the sole innervation to much of the male internal genitalia⁹, we compared the innervation of internal genitalia in fertile control and infertile $P52a/fu^MIR$ males (Fig. 3d, e). We found no differences in either the patterns of innervation by, or serotonergic differentiation of these neurons. These findings suggest that the infertility of $P52a/fu^MIR$ males stems from a defect in a median bundle function necessary for the proper transfer of sperm and seminal fluids (Fig. 4b), which is known to occur sequentially in this species¹⁶. Consistent with this notion, after copulation with $P52a/fu^MIR$ males, the presence of sperm and a mating plug were rarely detected in female internal genitalia and never together, whereas they were consistently seen in control matings (Supplementary Fig. 2).

Whereas the courtship behaviours of $P52a/fu^MIR$ males represent striking abnormalities in the *D. melanogaster* courtship ritual, it is worth noting that these aberrant mating strategies are similar to the normal strategies in other insect species. Numerous fly species (such as *Ephemeroptera* and *Fannia canicularis*)^{17,18} with aerial mating behaviour begin to track females almost immediately after encountering them (or indeed any object with comparable size and movement), and such tracking is immediately followed by attempts at copulation. Whereas multi-male courtships are abnormal for *D. melanogaster*, they are the norm in various other insect species, such as the solitary honeybee (*Anthophora plumipes*)¹⁹. The fact that the manipulation of a small set of Fru^M neurons in *D. melanogaster* generates behaviours resembling those of other species raises the

possibility that the CNS circuitry underlying courtship behaviours is common to many species and that ethological differences arise from relatively subtle variations in such a common ancestral circuit. For example, in *D. melanogaster* such changes may allow for

Table 1 Courtship and behavioural differences between $P52a/fu^MIR$ and control males

	<i>P52a/GFP</i>	<i>P52a/fu^{MIR}</i>
Courtship assays*		
Latency to initiation	94 ± 8 s	8 ± 1 s
Initiation to attempted copulation	111 ± 20 s	4 ± 0 s
Orienting	17/20	Not observed (0/20)
Tapping	15/20	Not observed (0/20)
Wing extension	20/20	20/20
Proboscis extension	19/20	20/20
Attempted copulation	17/20	20/20
Fertility	15/15	0/15
Short-term locomotor assays†		
Average line crossings (10 males; 4 min)	52 ± 10 s	52 ± 5 s
Multiple-male courtship assays‡		
Multiple-male courtship (<5 min) (2 males)	(2)/15	15/15
Male pair courtship latency to initiation (first male)	80 ± 8 s	10 ± 2 s
Male pair courtship latency to initiation (second male)	6:57 ± 3:27 min:s	12 ± 5 s

*Multiple courtship aberrations occur when Fru^M expression is inhibited in median bundle neurons. These include: (1) decreased courtship latency; (2) the absence of orienting and tapping; (3) a compressed courtship ritual; and (4) sterility. Values in bold indicate behaviours late in courtship that often occur simultaneously in $P52a/fu^MIR$ males.

†Short-term locomotor assays (4 min) show no differences in general activity between $P52a/fu^MIR$ and control males ($P = 0.98$).

‡Multiple-male courtship assays reveal that multiple $P52a/fu^MIR$ males invariably simultaneously court and attempt copulation with a single female. The non-courting male in pairs of control males only rarely displays courtship behaviour towards a female (2/15 times) within the first 5 min. Finally, there are differences in the courtship latencies of first and second males in pairs of control males when compared with pairs of $P52a/fu^MIR$ males.

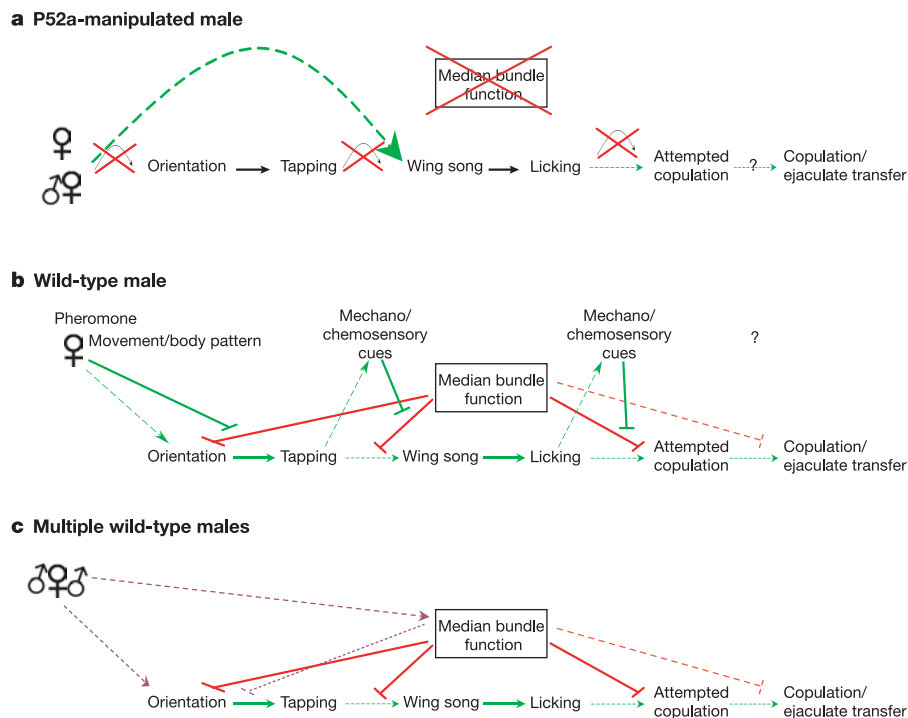


Figure 4 A functional schematic of putative median bundle roles during the *Drosophila* courtship ritual. **a**, Aberrant courtship behaviour in $P52a/fu^MIR$ males. In the absence of Fru^M -mediated median bundle function, elicitation of rapid-onset courtship by single females (green) or females in the presence of other males (purple, see **c**) bypasses initial steps (orientation and tapping) and rapidly progresses to attempted copulation. **b**, In the wild type, proper median bundle function is required for the staging of sequential courtship components and for proper transfer of seminal contents. Tonic inhibitions (red) of later behaviours are dependent upon Fru^M -expressing median bundle neurons. Disinhibitions of behaviours (green) are elicited in response to specific sensory cues.

Recognition of females initiates orienting, tracking and tapping behaviours by means of positively-acting sensory information. Additional cues may relieve the median-bundle-mediated inhibition of rapid courtship initiation and permit progression to tapping. Similar disinhibition following successful tapping allows wing song and licking, and successful licking disinhibits attempted copulation. Median bundle function is additionally needed for successful insemination (dashed red line). **c**, In wild-type situations involving multiple males, median bundle neurons function in the inhibition of courtship initiation by a second male when a female is already being courted (dashed purple line).

social dynamics appropriate to terrestrial environments where their courtship normally occurs²⁰.

How do median bundle neurons function to allow the appropriate coordination of the component behaviours of the courtship ritual? Clonal analysis of the morphology of these neurons (see Methods) reveals local dendritic elaborations in the suboesophageal ganglion, consistent with functional data suggesting that median bundle neurons may integrate information from multiple sensory modalities whose axons project to this region^{21,22} (Supplementary Fig. 3). Such connections would place median bundle neurons in an ideal position to modulate the initiation of component behaviours in response to appropriate, sequential, sensory cues. The projection of median bundle neurons directly onto the dorsal protocerebrum, a region generally implicated in higher-order processing, is also consistent with a role in coordinating/processing information from multiple modalities²¹.

The question then remains as to how male-specific Fru^M expression alters neuronal function. Examination of median bundle neurons in males and females labelled by *P52a-GAL4* fails to reveal any gross morphological differences between the sexes (data not shown). The existence of homologous neurons in females suggests that in this instance, Fru^M functions to modify neurons present in both sexes for male-specific functions. Whether Fru^M specifies changes in neuron morphology and/or physiology will no doubt illuminate the molecular mechanisms underlying the development of such circuitry, and may provide insights into how local development and changes in distinct subsets of neurons are then integrated and coordinated over evolutionary time to form complex learning and behavioural programs. □

Methods

Drosophila strains and cultures

The *P52a-GAL4*, *UAS-LacZ* line was provided by J. Nambu¹², and the *P52a-GAL4* driver was isolated and maintained independently. The Stinger 5 *UAS-GFP* lines was provided by S. Barolo. The *UAS-iraF* and *UAS-GFP* lines were obtained from the Bloomington *Drosophila* Stock Center. The *UAS-fru^MIR* line was generated as described below. Stocks were maintained on standard sugar media at 25 °C. Crosses were performed and progeny maintained at 29 °C. Males used in behavioural assays were collected 0–6 h after eclosion and raised in isolation on a 12 h light/dark cycle for 3–5 days before testing.

UAS-fru^MIR transgenic line

A pUAST-based construct designed to generate a hairpin duplex targeting amino acids 3–101 of the male-specific Fru^M N terminus was made as described²³ with 5' *EcoRI* and 3' *KpnI* sites. Transgenic lines were generated by P-element-mediated germ line transformation. The line used in these experiments contains two insertions of the *UAS-fru^MIR* construct.

CNS dissection, immunohistochemistry and imaging

Males at 0–12 h after eclosion were dissected, fixed and stained using rat polyclonal anti-Fru^M antisera (1:250) and Cy3-conjugated goat anti-rat IgG (1:1000; Jackson ImmunoResearch) as described¹¹. Stacks of optical sections, usually at 1 µm spacing, were obtained with a Bio-Rad MRC 1024 confocal microscope, using the Laser Sharp program, then processed with NIH Image and Adobe Photoshop.

Courtship assays and analysis

Courtship assays were performed in a similar way to established protocol¹⁴. Males were entrained in isolation for 3–5 days after eclosion and then single or multiple males were presented with a 1–2-day-old Canton S virgin. Flies were videotaped for 20 min or until copulation occurred, then courtship behaviour was analysed.

Locomotor assays and analysis

Assays were performed on ten males for each genotype and data were then averaged. Short-term locomotor assays were performed by placing a single male in a 35 mm Petri dish and counting the number of midline crossings in a 4-min-period. A standard t-test was performed to determine the significance of differences. Intermediate and long-term assays were performed as previously described¹⁴ using the *Drosophila* Activity Monitoring System (Trikinetics). Intermediate assays were performed for 4 h with 1-min-bins, and long-term assays were performed for 3 days, with food at one end of the monitor tube sealed with parafilm, with 30 min bins. Data were analysed using a non-parametric permutation test on all permutations to determine significance of differences.

Controls and other lines tested

In addition to the controls shown (Fig. 3a), other *GAL4* enhancer trap lines (c271, c552), also expressed in the median bundle, and other Fru^M-expressing neurons produced

similar decreases in courtship latency in *GAL4/fru^MIR* males, although this phenotype was less reproducible in c552 (data not shown). However, owing to overlap with other areas of Fru^M expression in the CNS, they were not used for further analysis. In experiments to examine the roles of other groups of Fru^M neurons in male courtship behaviour, using over 800 independent *GAL4* enhancer-trap insertion lines driving expression of *UAS-fru^MIR*, <1% of lines produced decreases in courtship latency similar to those seen in *GAL4/fru^MIR* males.

Male genital dissection, analysis and immunohistochemistry

Male genitalia were dissected from 3–5-day-old adults and either examined for motile sperm or fixed and stained with monoclonal antibody 22C10 (1:20; DSHB) and rabbit polyclonal anti-5HT antisera (1:800; Immunostar) and secondary antibodies Cy3-conjugated goat anti-mouse (1:1000; Jackson ImmunoResearch) and Oregon Green 488 goat anti-rabbit IgG (1:1500; Molecular Probes).

Matings and female genital dissections

Females were mated to *P52a/fru^MIR* males bearing Don Juan–GFP fusion-labelled sperm²⁴. Female internal genitalia were dissected within 45 min after the end of copulation and examined using brightfield and fluorescent microscopy.

Clonal analysis

Clones were generated as previously described²⁵ by delivering a 5 min 37 °C heat shock to 6–30-h-old embryos containing the *hsFLP*, *P52a-GAL4* and *UAS > CD2, y + > CD8-GFP* and then raised to adulthood at 18 °C. CNS dissection and immunohistochemistry was performed as described with rat anti-mouse CD8 (1:100; Caltag Laboratories) and Alexafluor 488-conjugated goat anti-rat (1:200; Molecular Probes).

Received 13 January; accepted 4 June 2004; doi:10.1038/nature02713.

1. Baker, B. S., Taylor, B. J. & Hall, J. C. Are complex behaviors specified by dedicated regulatory genes? Reasoning from *Drosophila*. *Cell* **105**, 13–24 (2001).
2. Emmons, S. W. & Lipton, J. Genetic basis of male sexual behavior. *J. Neurobiol.* **54**, 93–110 (2003).
3. Greenspan, R. J. & Ferveur, J. E. Courtship in *Drosophila*. *Annu. Rev. Genet.* **34**, 205–232 (2000).
4. Hall, J. C. The mating of a fly. *Science* **264**, 1702–1714 (1994).
5. Manning, A. The sexual behaviour of two sibling *Drosophila* species. *Behaviour* **15**, 123–145 (1959).
6. Ryner, L. C. *et al.* Control of male sexual behavior and sexual orientation in *Drosophila* by the fruitless gene. *Cell* **87**, 1079–1089 (1996).
7. Ito, H. *et al.* Sexual orientation in *Drosophila* is altered by the satori mutation in the sex-determination gene *fruitless* that encodes a zinc finger protein with a BTB domain. *Proc. Natl Acad. Sci. USA* **93**, 9687–9692 (1996).
8. Anand, A. *et al.* Molecular genetic dissection of the sex-specific and vital functions of the *Drosophila melanogaster* sex determination gene *fruitless*. *Genetics* **158**, 1569–1595 (2001).
9. Lee, G., Villella, A., Taylor, B. J. & Hall, J. C. New reproductive anomalies in fruitless-mutant *Drosophila* males: extreme lengthening of mating durations and infertility correlated with defective serotonergic innervation of reproductive organs. *J. Neurobiol.* **47**, 121–149 (2001).
10. Goodwin, S. F. *et al.* Aberrant splicing and altered spatial expression patterns in fruitless mutants of *Drosophila melanogaster*. *Genetics* **154**, 725–745 (2000).
11. Lee, G. *et al.* Spatial, temporal, and sexually dimorphic expression patterns of the fruitless gene in the *Drosophila* central nervous system. *J. Neurobiol.* **43**, 404–426 (2000).
12. Melnattur, K., Rawson, E. & Nambu, J. R. P[52A–GAL4] is an insertion in the *Drosophila* GP150 gene. *Genesis* **34**, 29–33 (2002).
13. Gaines, P., Tompkins, L., Woodard, C. T. & Carlson, J. R. *quick-to-court*, a *Drosophila* mutant with elevated levels of sexual behavior, is defective in a predicted coiled-coil protein. *Genetics* **154**, 1627–1637 (2000).
14. Villella, A. *et al.* Extended reproductive roles of the fruitless gene in *Drosophila melanogaster* revealed by behavioral analysis of new *fru* mutants. *Genetics* **147**, 1107–1130 (1997).
15. Bray, S. & Amrein, H. A putative *Drosophila* pheromone receptor expressed in male-specific taste neurons is required for efficient courtship. *Neuron* **39**, 1019–1029 (2003).
16. Alonso-Pimentel, H., Tolbert, L. P. & Heed, W. B. Ultrastructural examination of the insemination reaction in *Drosophila*. *Cell Tissue Res.* **275**, 467–479 (1994).
17. Spieth, H. T. Studies on the biology of the *Ephemeroptera*. II. The nuptial flight. *J. NY Entomol. Soc.* **48**, 379–390 (1940).
18. Land, M. F. & Collett, T. S. Chasing behavior of houseflies (*Fannia canicularis*). *J. Comp. Physiol.* **89**, 331–357 (1974).
19. Stone, G. N., Loder, P. M. J. & Blackburn, T. M. Foraging and courtship behaviour in males of the solitary bee *anthophora plumipes* (Hymenoptera, Anthophoridae)—Thermal physiology and the roles of body size. *Ecol. Entomol.* **20**, 169–183 (1995).
20. Spieth, H. T. Courtship behavior in *Drosophila*. *Annu. Rev. Entomol.* **19**, 385–405 (1974).
21. Zars, T., Wolf, R., Davis, R. & Heisenberg, M. Tissue-specific expression of a type I adenylyl cyclase rescues the rutabaga mutant memory defect: in search of the engram. *Learn. Mem.* **7**, 18–31 (2000).
22. Shanbhag, S. & Singh, R. N. Functional implications of the projections of neurons from individual labellar sensillum of *Drosophila melanogaster* as revealed by neuronal marker horseradish peroxidase. *Cell Tissue Res.* **267**, 273–282 (1992).
23. Billuart, P., Winter, C. G., Maresh, A., Zhao, X. & Luo, L. Regulating axon branch stability: the role of p190 RhoGAP in repressing a retraction signaling pathway. *Cell* **107**, 195–207 (2001).
24. Santel, A., Winhauer, T., Blumer, N. & Renkawitz-Pohl, R. The *Drosophila* *don juan* (*dj*) gene encodes a novel sperm specific protein component characterized by an unusual domain of a repetitive amino acid motif. *Mech. Dev.* **64**, 19–30 (1997).
25. Wong, A. M., Wang, J. W. & Axel, R. Spatial representation of the glomerular map in the *Drosophila* protocerebrum. *Cell* **109**, 229–241 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The authors thank R. Fernald, J. Hall, L. Luo, S. McConnell, B. Taylor, N. Anandasabapathy, M. Arbeitman, E. Marin and members of the Baker laboratory for comments and discussions; M. Siegal for help with statistical analysis; J. Nambu and K. Melnattur for the *P52a-GAL4* line; S. Barolo for the *UAS-GFPnls* line; E. Reynaud for the rat anti-Fru^M antiserum; G. Bohm for the preparation of culture media and fly food. This work was supported by the Medical Scientist Training Program (D.S.M.) and by an NIH grant to B.S.B.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.S.B. (bbaker@cmgm.stanford.edu).

Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1

David M. Sayah¹, Elena Sokolskaja¹, Lionel Berthoux¹ & Jeremy Luban^{1,2}

¹Department of Microbiology and ²Department of Medicine, Columbia University, College of Physicians and Surgeons, 701 West 168th Street, HHSC 1502 New York, New York 10032, USA

In Old World primates, TRIM5- α confers a potent block to human immunodeficiency virus type 1 (HIV-1) infection that acts after virus entry into cells^{1–5}. Cyclophilin A (CypA) binding to viral capsid protects HIV-1 from a similar activity in human cells^{4,6–8}. Among New World primates, only owl monkeys exhibit post-entry restriction of HIV-1 (ref. 1). Paradoxically, the barrier to HIV-1 in owl monkey cells is released by capsid mutants or drugs that disrupt capsid interaction with CypA⁴. Here we show that knockdown of owl monkey CypA by RNA interference (RNAi) correlates with suppression of anti-HIV-1 activity. However, reintroduction of CypA protein to RNAi-treated cells did not restore antiviral activity. A search for additional RNAi targets unearthed *TRIMCyp*, an RNAi-responsive messenger RNA encoding a TRIM5–CypA fusion protein. *TRIMCyp* accounts for post-entry restriction of HIV-1 in owl monkeys and blocks HIV-1 infection when transferred to otherwise infectable human or rat cells. It seems that *TRIMCyp* arose after the divergence of New and Old World primates when a LINE-1 retrotransposon catalysed the insertion of a *CypA* complementary DNA into the *TRIM5* locus. This is the first vertebrate example of a chimaeric gene generated by this mechanism of exon shuffling.

Post-entry restriction of HIV-1 infection is common among Old World monkeys, but owl monkeys are unique among New World primates in exhibiting this phenotype¹. *Aotus trivirgatus* owl monkey kidney cells (OMK) restrict HIV-1 infection, but are permissive for simian immunodeficiency virus (SIV) infection¹. HIV-1 restriction in OMK cells is completely abrogated when the interaction between HIV-1 capsid and the cellular protein cyclophilin A (CypA) is disrupted⁴, either by mutations altering capsid or by treatment of target cells with the cyclophilin-binding drug cyclosporin A (CsA). This phenotype is the opposite of that seen in most human cells where the capsid–CypA interaction is required for efficient HIV-1 replication^{4,6–9}.

The paradoxical response to CsA in OMK cells was investigated further using two other drugs: Melle⁴-CsA, a non-immunosuppressive analogue⁹, and sanglifehrin, a structurally unrelated compound that also binds cyclophilin¹⁰. As with CsA, treatment of target cells with these compounds permits HIV-1 to infect OMK cells at an

efficiency similar to that of SIV (Fig. 1a). Peripheral blood mononuclear cells (PBMC) from a different owl monkey species, *Aotus nancymae*, show the same restriction phenotype with respect to SIV, HIV-1 and CsA (Fig. 1b). Identical results were obtained with PBMC from a second animal (data not shown).

CsA, Melle⁴-CsA and sanglifehrin bind cyclophilin family members, but not exclusively CypA. To examine the specific role of CypA, we generated stable OMK cell lines with CypA knockdown by RNAi. Of six CypA-specific small hairpin RNAs (shRNAs), three decreased CypA expression (Fig. 1c). Those shRNA constructs that decreased CypA expression abrogated HIV-1 restriction to a corresponding degree (Fig. 1d).

To determine whether disruption of HIV-1 restriction was due to CypA knockdown, the OMK knockdown cell line with the largest decrease in CypA expression and HIV-1 restriction (OMK_{MH-CypA-147}) was transfected with non-targetable *CypA* cDNAs (*ntCypA*) bearing silent mutations to make them resistant to the RNAi. A plasmid encoding cell surface H-2K^K was cotransfected so that transfected cells could be enriched using antibodies conjugated to magnetic particles. Although cells selected in this manner were fully restored for CypA expression, they remained deficient for HIV-1 restriction (Fig. 1e). We attempted to restore

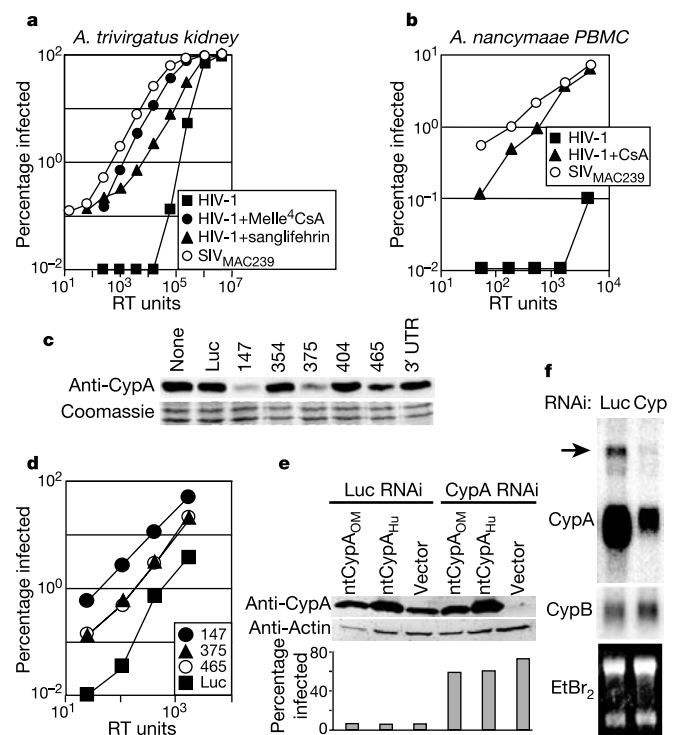


Figure 1 A CypA homologue is required for owl monkey restriction of HIV-1. **a**, *Aotus trivirgatus* OMK cells (**a**) or *Aotus nancymae* peripheral blood mononuclear cells (**b**) were infected with GFP-transducing virions in the presence of CypA-binding drugs. GFP-positive cells were counted by flow cytometry. RT units, relative amount of virus added, as measured by reverse transcriptase activity. **c**, **d**, OMK cells were transfected with retroviruses delivering shRNAs targeting CypA mRNA at the indicated nucleotide positions. Lysates were immunoblotted (**c**) and cells were challenged with HIV-1/GFP (**d**). After antibody detection, the membrane was Coomassie-stained and a representative section is shown as a loading control. Luc is the control shRNA targeting luciferase. **e**, OMK cells transfected with shRNA-147 were selected after transfection with non-targetable-CypA expression vectors (human or owl monkey) and immunoblotted (top), or infected with HIV-1/GFP (bottom). **f**, Northern blot of total cytoplasmic RNA from shRNA-147- or shRNA-Luc-treated OMK cells, probed with *CypA* and *CypB* cDNA. The arrow indicates a ~2-kb, RNAi-responsive mRNA that hybridizes to *CypA*. EtBr₂, ethidium bromide.

CypA expression in OMK_{MH-CypA-147} cells several times using different methods, including retroviral transduction of *ntCypA* cDNA. In all cases, reconstitution of CypA expression did not restore HIV-1 restriction activity.

Off-target spread of RNAi has been reported¹¹, but this seemed an unlikely explanation for our inability to restore restriction to OMK_{MH-CypA-147} because CypA knockdown with three different RNAi target sequences was associated with loss of restriction. A more plausible explanation seemed to be knockdown of an unknown mRNA homologous to CypA. Multiple screens of OMK cDNA by hybridization or polymerase chain reaction (PCR) yielded only bona fide CypA cDNA. Northern analysis was performed to look for transcripts other than CypA that were decreased by RNAi in these cells. Total cytoplasmic RNA from OMK_{MH-CypA-147}, or from control OMK cells treated with RNAi specific for luciferase, was hybridized with CypA coding sequence (Fig. 1f). The dominant CypA transcript (about 750 nucleotides) was significantly reduced in OMK_{MH-CypA-147}. Another transcript of approximately 2 kb hybridized to the CypA probe and was also decreased by RNAi against CypA.

To identify the larger transcript detected by northern blot, a size-selected OMK cDNA library (1,700 to 2,500 base pairs, bp) was screened by colony hybridization with a CypA probe. Inserts from four positive colonies were sequenced. Two contained in-frame fusions between TRIM5 (exons 1 to 7) and a complete CypA cDNA. This TRIMCyp cDNA was predicted to encode a 54-kD protein chimera consisting of the amino-terminal 299 amino acids of TRIM5 and the 165-amino-acid CypA protein linked by 11 amino acids encoded by the CypA 5' untranslated region (UTR) (Fig. 2a). 293T cells transfected with a TRIMCyp cDNA expression vector produced a protein of the expected size that was reactive with both anti-TRIM5 and anti-CypA antibodies (Fig. 2b).

TRIMCyp mRNA was detectable in OMK cells and in PBMC from two *Aotus nancymaae* monkeys (Fig. 2c, d), but not in the human cell lines Jurkat or HeLa, or in rhesus macaque FRhK4 cells (Fig. 2c). Quantitative reverse transcription (RT)-PCR was used to demonstrate that TRIMCyp expression in the CypA knockdown cells used in Fig. 1e, f was decreased tenfold (Fig. 2e).

To determine whether TRIMCyp was sufficient to restore HIV-1 restriction to CypA knockdown OMK cells, OMK_{MH-CypA-147} were transduced with a pMIG retroviral vector delivering a TRIMCyp cDNA bearing silent mutations that render it resistant to RNAi (*ntTRIMCyp*). pMIG expresses cDNAs fused to an internal ribosome entry site (IRES)-green fluorescent protein (GFP) cassette, permitting identification of transduced cells by GFP expression. pMIG-transduced OMK_{MH-CypA-147} cells were assayed for their ability to restrict an HIV-1 vector that delivers the DsRed fluorescent protein. OMK_{MH-CypA-147} cells transduced with pMIG-*ntTRIMCyp* restricted HIV-1, while those transduced with empty vector, or vector delivering *ntCypA*_{OM}, did not (Fig. 2f). Loss of restriction in OMK cells caused by RNAi against CypA (Fig. 1c, d) could therefore be attributed to TRIMCyp knockdown.

To see if transfer of TRIMCyp confers post-entry restriction of HIV-1 to cells derived from other species, we generated stable human TE671 cell lines expressing TRIMCyp, CypA_{OM}, or TRIMStop, a construct in which TRIMCyp has been truncated with a stop codon replacing the first codon of the CypA domain. TE671 cells expressing TRIMCyp were highly resistant to HIV-1 infection, while expression of TRIMStop or CypA_{OM} had no effect on HIV-1 titre (Fig. 3a). As expected, SIV_{MAC239} infection of TE671 cells was unaffected by TRIMCyp (Fig. 3b). HIV-1 restriction mediated by TRIMCyp was bypassed by treating target cells with CsA during infection, or by infection with HIV-1 bearing the G89V capsid mutation that prevents CypA binding (Fig. 3c). HeLa cells transduced with pMIG-TRIMCyp exhibited the same restriction (Fig. 3d). These data demonstrate that HIV-1 restriction by TRIMCyp is capsid-specific and cell-type independent. The CypA domain

of TRIMCyp is essential for restriction and disruption of CypA-capsid binding abrogates TRIMCyp-mediated restriction.

Sequencing of OMK genomic DNA, and comparison with human sequence, revealed a complete CypA cDNA between TRIM5 exons 7 and 8 (Fig. 4a). We did not detect an owl monkey TRIM5 allele lacking the CypA insertion. The inserted CypA cDNA is 95% identical to the bona fide owl monkey CypA cDNA, but only 88% identical to human or rhesus macaque CypA, suggesting that CypA insertion into TRIM5 occurred after the divergence of New and Old World primates. Because other New World monkeys lack post-entry HIV-1 restriction activity¹, the CypA insertion appears to be unique to owl monkeys.

The CypA insertion bears the hallmarks of LINE-1 (L1)-mediated

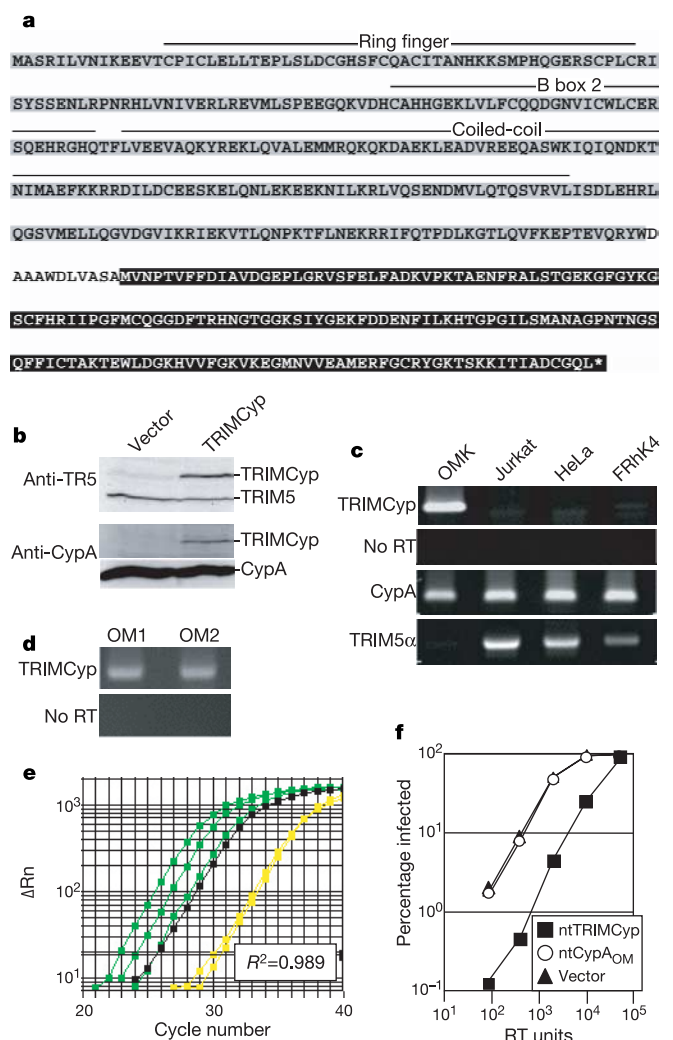


Figure 2 Owl monkey cells express a TRIM5-CypA fusion protein that blocks HIV-1 infection. **a**, Predicted amino-acid sequence of 54-kD TRIMCyp protein. TRIM5 segment highlighted in grey with domains indicated above; CypA segment highlighted in black. **b**, Immunoblot of 293T cells transfected with TRIMCyp expression vector, probed for TRIM5 and CypA. **c**, **d**, RT-PCR for TRIMCyp, CypA or TRIM5-α in OMK, human HeLa and Jurkat, and macaque FRhK4 cells (**c**), and PBMC from two *Aotus nancymaae* monkeys (**d**). **e**, Amplification plot of real time RT-PCR. The green curves show undiluted and threefold serial dilutions of OMK_{MH-Luc} cDNA; the black curve is undiluted OMK_{MH-CypA-147} cDNA; the yellow curves are 'no RT' controls. Rn, normalized reporter signal, R^2 , the correlation coefficient for the diluted samples. GAPDH amplification curves (not shown) were identical between samples. **f**, OMK_{MH-CypA-147} cells transduced with the indicated RNAi-resistant cDNAs and infected with HIV-1.

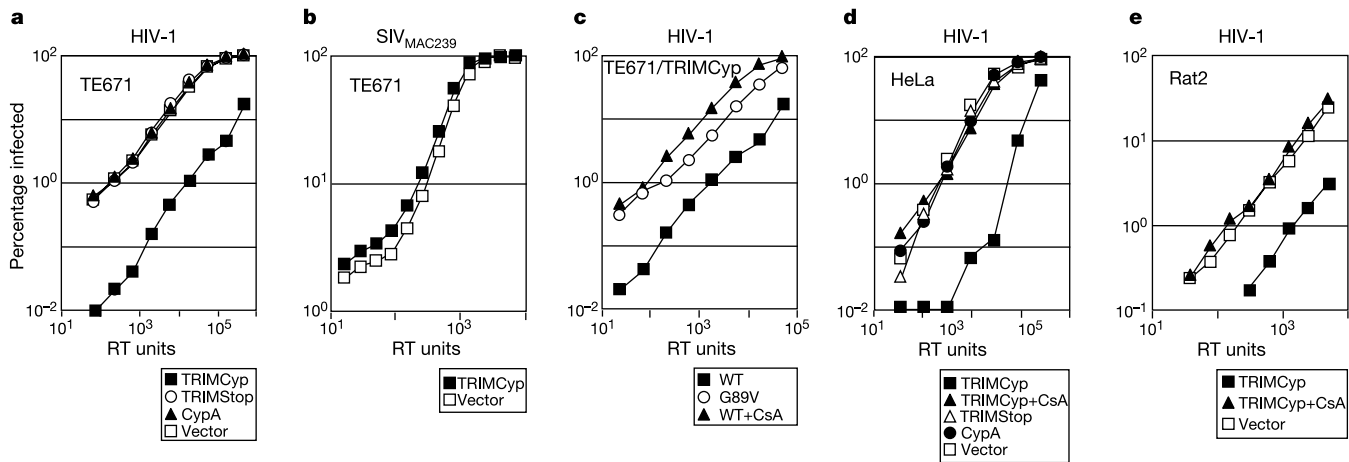


Figure 3 HIV-1 is restricted in human or rat cells transduced with *TRIMCyp*. TE671 cells transduced with the indicated cDNAs were infected with HIV-1 (**a**) or SIV_{MAC239} (**b**). **c**, *TRIMCyp*-transduced TE671 cells were infected with HIV-1 $\pm$ CsA or with HIV-1 bearing

the capsid G89V mutation. WT, wild type. HeLa cells (**d**) or Rat2 cells (**e**) transduced with the indicated cDNAs were infected with HIV-1. Results shown are typical of those obtained in three independent experiments.

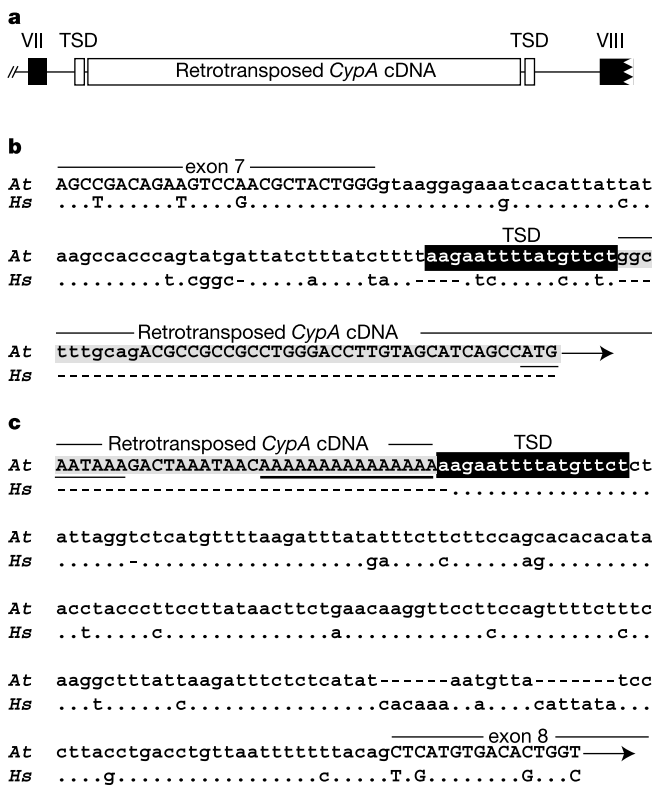


Figure 4 *TRIMCyp* arose from retrotransposition of *CypA* cDNA into *TRIM5*. **a**, Part of the owl monkey *TRIM5* locus showing a complete, processed *CypA* cDNA inserted after exon 7. **b**, **c**, *TRIM5* genomic sequence flanking the 5' end (**b**) or the 3' end (**c**) of the *CypA* insertion aligned with homologous human sequence. *TRIMCyp* exons and the region homologous to human exon 8 are in capital letters, introns in lower case, and the *CypA* cDNA insertion highlighted in grey. The 'natural' *CypA* start codon, the *CypA/TRIMCyp* poly-adenylation signal and poly-A tail are underlined. The insertion is flanked by a 16-bp target site duplication (TSD), highlighted in black. At, *Aotus trivirgatus*. Hs, *Homo sapiens*.

retrotransposition¹². These include flanking 16-bp direct repeats consistent with target site duplication, a processed cDNA devoid of introns and accompanied by a poly-A tail, and insertion at a consensus, A-rich L1 endonuclease recognition site¹³ (Fig. 4b, c).

Mammalian genomes possess many reverse-transcribed *CypA* pseudogenes which, by definition, are defective¹⁴. *TRIMCyp* is unusual in that a complete, processed *CypA* cDNA has generated a new *CypA* exon that is expressed and spliced to *TRIM5* exon 7 to encode a novel chimaeric protein.

TRIMCyp's mechanism of action may be hinted at by its structure and by recent experiments with *TRIM5* homologues from other species. The *TRIM5* ring-finger domain confers E3 ubiquitin ligase activity¹⁵, suggesting that *TRIMCyp* attaches ubiquitin to incoming HIV-1 virion proteins. Human cells possess a *TRIM5* homologue and HIV-1 infection of these cells is stimulated by proteasome inhibitors¹⁶. We were unable to detect an effect of these drugs on HIV-1 restriction of OMK cells, although these cells are resistant to other drugs, including antibiotics commonly used for selection of mammalian clones. *TRIM5* possesses a coiled-coil domain which causes the protein to aggregate and form cytoplasmic bodies^{15,17}. *TRIMCyp* forms similar structures and associates with human *TRIM5- α* in the yeast two-hybrid system (data not shown). Though *TRIMCyp* is functional in non-primate cells (rat fibroblasts, Fig. 3e), restriction may require recruitment of other *TRIM* family members.

The *TRIMCyp* carboxy terminus, which is required for restriction (Fig. 3a, d), bears a complete *CypA* protein that probably functions autonomously to recognize HIV-1 capsid⁶. This would not be surprising because cyclophilins frequently exist as discrete domains within complex proteins¹⁸. In rhesus macaques, *TRIM5- α* accounts for a potent HIV-1 restriction, and the C-terminal B30.2 (SPRY) domain that distinguishes the α -isoform of *TRIM5* is essential for this activity⁵. RT-PCR with primers based on the owl monkey genomic sequence failed to detect an owl monkey *TRIM5- α* transcript (one containing the exon 8-encoded B30.2 domain instead of the *CypA* domain) (Fig. 2c). Because the *CypA* domain of *TRIMCyp* substitutes for the B30.2 domain present in macaque *TRIM5- α* , it is likely that the C terminus of *TRIM5- α* is also responsible for recognition of incoming viral capsids.

Since the discovery of transposons over 50 years ago¹⁹, it has become clear that these elements play a fundamental role in evolution. L1 elements are the most abundant autonomous retrotransposons in the human genome and are believed to function *in trans* to retrotranspose *Alu* elements and processed pseudogenes²⁰⁻²². Consistent with the exon shuffling model²³, L1 elements have been shown experimentally to transfer non-L1 sequences into existing genes to generate new gene chimaeras²⁴. Outside the

laboratory, a retrotransposed SVA element forms the last exon of a leptin receptor²⁵ and one *ATM* exon retrotransposed into a new genomic location²⁶. The only previous report of a retrotransposed complete mRNA generating a new exon in a previously existing gene is *jingwei* in *Drosophila melanogaster*²⁷. *TRIMCyp* is the first example of the genesis of such a chimaeric gene in vertebrates. Interestingly, among primates, owl monkeys have relatively high rates of *Alu* amplification²⁸, an indication of increased L1-mediated transposition in these animals. The generation of *TRIMCyp* probably represents one example of a process that has played a role in the evolution of many genes. □

Methods

Plasmids

Most constructs were previously described⁴. CSRW, an HIV-1 vector derived from CSGW, bears DsRed (Clontech). p8.9NΔSB is an HIV-1 packaging vector based on pCMVΔR8.91. pMACS-KK.II (Miltenyi Biotec) encodes truncated murine cell-surface major histocompatibility complex (MHC) I H-2K^K.

Oligonucleotides encoding shRNAs targeting CypA (Supplementary Table 1, oligos 9–20) or firefly luciferase (oligos 21 and 22) were ligated into pSUPER²⁹. The shRNA expression cassette was subcloned into pMSCVΔU3³⁰ to generate pMH.

ntTRIMCyp was generated by PCR using oligos 24 and 25, and 23 and 26. *TRIMStop* replaces the CypA start codon with a stop codon (oligos 25 and 29). *ntCypA_{OM}* and *ntCypA_{Hu}* are CypA cDNAs resistant to pMH–CypA147 that were generated by PCR using oligos 23–27. *ntTRIMCyp*, *TRIMStop* and *ntCypA* were cloned into pcDNA3.1, pMIG and pLPCX.

Northern Blots

Total cytoplasmic RNA (RNeasy, Qiagen) was fractionated by standard protocols using formaldehyde agarose gel electrophoresis, transferred to Hybond-N+ nylon membranes (Amersham), and fixed by ultraviolet crosslinking (Stratalinker, Stratagene). Probes were owl monkey *CypA* or human *CypB* coding sequence, ³²P-labelled by random priming (RediPrimeII, Amersham). Blots were hybridized in Rapid-hyb (Amersham), and visualized on a PhosphorImager (Molecular Devices).

RT–PCR and real time RT–PCR

Total RNA from cultured cells or PBMC (RNeasy, Qiagen) was reverse-transcribed into cDNA by random priming (Superscript II, Invitrogen). Primers 1 and 2 were used to amplify *TRIMCyp* cDNA, primers 5 and 6 to amplify *CypA*, and primers 30 and 31 to amplify *TRIM5-α*. Real-time RT–PCR was performed as previously described³⁰ with product detection using SYBR Green (Molecular Probes). Primers 1 and 2 were used to amplify *TRIMCyp*. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) control curves (primers 7 and 8) were identical between samples in the experiments presented here. Cycling conditions were 40 cycles: 94 °C, 15 s; 55 °C, 15 s; 72 °C, 30 s.

cDNA library screen

OMK cDNA was size-selected (1,700–2,500 bp) by agarose gel electrophoresis and cloned into pBluescriptII-KS. 100,000 colonies were screened by hybridization with a *CypA* probe. OMK *CypA* cDNA was cloned similarly but without size selection.

Genomic PCR

OMK genomic DNA was amplified by PCR using oligos 1 and 2, oligos 3 and 4, or oligos 32 and 33 (Supplementary Table 1). PCR products were purified (Qiaquick, Qiagen) and sequenced.

Cell lines and PBMCs

OMK, 293T, TE671, HeLa and Rat2 cells were obtained from ATCC. Owl monkey PBMC were isolated with Ficoll-Paque Plus (Pharmacia), stimulated with 2 μg ml^{−1} PHA (Sigma), and cultured in RPMI supplemented with 20% FBS, 20 U ml^{−1} recombinant human IL-2 (Boehringer Mannheim) and 2 mM L-glutamine.

Drugs

CsA (Bedford Laboratories) was prepared in DMSO at 10 mg ml^{−1}. Melle⁴–CsA and sanglifehrin (gifts from Novartis, Basel) were prepared in DMSO at 10 mg ml^{−1} and 10 mM, respectively. All drugs were diluted in tissue culture medium to 2.5 μM and added at the time of infection.

Immunoblots

Whole-cell extracts were fractionated by SDS–polyacrylamide gel electrophoresis (PAGE), transferred to PVDF membranes, and probed using rabbit anti-CypA (Biomol), goat anti-TRIM5 (Abcam), or goat anti-actin (Santa Cruz Biotechnology).

Viruses and retroviral vectors

Vectors and viruses were produced by transfection of 293T cells using lipofectamine 2000 (Invitrogen) and pseudotyped with VSV-G. pMH, pMIG and pLPCX vectors were produced by co-transfection of the vector plasmid, pCL-Eco and pMD.G using a 5:5:1 ratio of the three plasmids in each transfection. HIV-DsRed and HIV-GFP vectors were produced in 293T cells by co-transfection of CSGW or CSRW with p8.9NΔSB (wild type

or G89V mutant) and pMD.G, also at 5:5:1 ratio. HIV/Env-/GFP and SIV/Env-/GFP viruses were produced by cotransfection of the wild type or G89V HIV/Env-/GFP plasmids with pMD.G at a 10:1 ratio. All vectors and viruses were harvested 48 h post-transfection, filtered (0.45-μm filter, Pall Acrodisc), aliquoted, and stored at −80 °C. Stocks were normalized by reverse transcriptase activity or p24 ELISA⁸.

OMK cell transfection and magnetic selection

5 × 10⁵ OMK cells per well in six-well plates were co-transfected with 1 μg pcDNA3.1 expression vector and 0.1 μg pMACS-K^K.II using 6 μl lipofectamine 2000. H-2K^K-positive cells were selected (MACSelect KK.II, Miltenyi Biotec) 24 h post-transfection, and either processed for western blot or infected with HIV-GFP.

Generation of stable cell lines

To generate stable RNAi knockdown lines, 1 ml of pMH vector-containing media was pelleted onto OMK cells (1,200g for 72 min) in 12-well plates. Infection was repeated 24 h later. *ntTRIMCyp*, *TRIMStop* and *ntCypA_{OM}* cDNAs were transduced with either pMIG (OMK and HeLa cells) or pLPCX (TE671 and Rat2 cells). pLPCX-infected cells were selected in 1 μg ml^{−1} puromycin.

Infection assays

Cells were assayed by flow cytometry 48 to 72 h post-infection to determine the percentage of cells that were GFP-positive (FL-1) or DsRed-positive (FL-2). Cells transduced with pMIG vectors and infected with HIV-DsRed were gated to limit the analysis to GFP-positive cells.

Received 24 May; accepted 22 June 2004; doi:10.1038/nature02777.

Published online 7 July 2004.

- Hofmann, W. *et al.* Species-specific, postentry barriers to primate immunodeficiency virus infection. *J. Virol.* **73**, 10020–10028 (1999).
- Cowan, S. *et al.* Cellular inhibitors with Fv1-like activity restrict human and simian immunodeficiency virus tropism. *Proc. Natl Acad. Sci. USA* **99**, 11914–11919 (2002).
- Besnier, C., Takeuchi, Y. & Towers, G. Restriction of lentivirus in monkeys. *Proc. Natl Acad. Sci. USA* **99**, 11920–11925 (2002).
- Towers, G. *et al.* Cyclophilin A modulates the sensitivity of HIV-1 to host restriction factors. *Nature Med.* **9**, 1138–1143 (2003).
- Stremlau, M. *et al.* The cytoplasmic body component TRIM5α restricts HIV-1 infection in Old World monkeys. *Nature* **427**, 848–853 (2004).
- Luban, J., Bossolt, K. L., Franke, E. K., Kalpana, G. V. & Goff, S. P. Human immunodeficiency virus type 1 Gag protein binds to cyclophilins A and B. *Cell* **73**, 1067–1078 (1993).
- Braaten, D. & Luban, J. Cyclophilin A regulates HIV-1 infectivity, as demonstrated by gene targeting in human T cells. *EMBO J.* **20**, 1300–1309 (2001).
- Franke, E. K., Yuan, H. E. & Luban, J. Specific incorporation of cyclophilin A into HIV-1 virions. *Nature* **372**, 359–362 (1994).
- Thali, M. *et al.* Functional association of cyclophilin A with HIV-1 virions. *Nature* **372**, 363–365 (1994).
- Fehr, T., Kallen, J., Oberer, L., Sanglier, J. J. & Schilling, W. Sanglifehrins A, B, C and D, novel cyclophilin-binding compounds isolated from *Streptomyces* sp. A92–308110. II. Structure elucidation, stereochemistry and physico-chemical properties. *J. Antibiot. (Tokyo)* **52**, 474–479 (1999).
- Jackson, A. L. *et al.* Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnol.* **21**, 635–637 (2003).
- Kazanian, H. H. Jr Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
- Feng, Q., Moran, J. V., Kazanian, H. H. Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916 (1996).
- Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13**, 2541–2558 (2003).
- Xu, L. *et al.* BTBD1 and BTBD2 colocalize to cytoplasmic bodies with the RBCC/tripartite motif protein, TRIM5β. *Exp. Cell Res.* **288**, 84–93 (2003).
- Schwartz, O., Marechal, V., Friguet, B., Arenzana-Seisdedos, F. & Heard, J. M. Antiviral activity of the proteasome on incoming human immunodeficiency virus type 1. *J. Virol.* **72**, 3845–3850 (1998).
- Reymond, A. *et al.* The tripartite motif family identifies cell compartments. *EMBO J.* **20**, 2140–2151 (2001).
- Wu, J., Matunis, M. J., Kraemer, D., Blobel, G. & Coutavas, E. Nup358, a cytoplasmically exposed nucleoporin with peptide repeats, Ran-GTP binding sites, zinc fingers, a cyclophilin A homologous domain, and a leucine-rich region. *J. Biol. Chem.* **270**, 14209–14213 (1995).
- McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl Acad. Sci. USA* **36**, 344–355 (1950).
- Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24**, 363–367 (2000).
- Wei, W. *et al.* Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell Biol.* **21**, 1429–1439 (2001).
- Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nature Genet.* **35**, 41–48 (2003).
- Gilbert, W. Why genes in pieces? *Nature* **271**, 501 (1978).
- Moran, J. V., DeBerardinis, R. J. & Kazanian, H. H. Jr Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).
- Damert, A., Lower, J. & Lower, R. Leptin receptor isoform 219.1: an example of protein evolution by LINE-1-mediated human-specific retrotransposition of a coding SVA element. *Mol. Biol. Evol.* **21**, 647–651 (2004).
- Ejima, Y. & Yang, L. Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. *Hum. Mol. Genet.* **12**, 1321–1328 (2003).
- Long, M. & Langley, C. H. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95 (1993).

28. Roy-Engel, A. M. *et al.* Non-traditional Alu evolution and primate genomic diversity. *J. Mol. Biol.* **316**, 1033–1040 (2002).
29. Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550–553 (2002).
30. Asmal, M. *et al.* Production of ribosome components in effector CD4+ T cells is accelerated by TCR stimulation and coordinated by ERK-MAPK. *Immunity* **19**, 535–548 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Bieniasz, S. Goff, M. Nicoletis, E. Phelps, S. Ribeiro, G. Towers and S. Valente for reagents. This work was supported by the NIH (J.L.), the Medical Scientist Training Program (D.M.S.), and AmFAR (L.B.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.L. (jl45@columbia.edu). TRIMCyp and CypA sequences have been deposited in GenBank (accession numbers AY646198, AY646199 and AY646200).

Cohesin relocation from sites of chromosomal loading to places of convergent transcription

Armelle Lengronne¹, Yuki Katou³, Saori Mori^{4,5}, Shihori Yokobayashi⁶, Gavin P. Kelly², Takehiko Itoh⁷, Yoshinori Watanabe^{6,8}, Katsuhiko Shirahige^{3,4*} & Frank Uhlmann^{1*}

¹Chromosome Segregation Laboratory, and ²Computational Genome Analysis Laboratory, Cancer Research UK London Research Institute, Lincoln's Inn Fields Laboratories, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

³Riken Genomic Science Center, Human Genome Research Group, Genome Informatics Team, 1-7-22-W417 Suehiro, Tsurumi, Yokohama City, Kanagawa, 230-0045, Japan

⁴Center for Biological Resources and Informatics, Division of Gene Research, and Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8501, Japan

⁵Science of Biological Supramolecular Systems, Graduate School of Integrated Science, Yokohama City University, 1-7-29 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan

⁶Institute of Molecular and Cellular Biosciences, University of Tokyo, Yayoi 1-1-1, Tokyo 113-0032, Japan

⁷Research Center for Advanced Science and Technology, Mitsubishi Research Institute Inc., 2-3-6 Ohtemachi, Chiyoda-ku, Tokyo, 100-8141, Japan

⁸SORST, Japan Science and Technology Agency, Yayoi 1-1-1, Tokyo 113-0032, Japan

* These authors contributed equally to this work

Sister chromatids, the products of eukaryotic DNA replication, are held together by the chromosomal cohesin complex after their synthesis. This allows the spindle in mitosis to recognize pairs of replication products for segregation into opposite directions^{1–6}. Cohesin forms large protein rings that may bind DNA strands by encircling them⁷, but the characterization of cohesin binding to chromosomes *in vivo* has remained vague. We have performed high resolution analysis of cohesin association along budding yeast chromosomes III–VI. Cohesin localizes almost exclusively between genes that are transcribed in converging directions. We find that active transcription positions cohesin at these sites, not the underlying DNA sequence. Cohesin is initially loaded onto chromosomes at separate places, marked by the Scc2/Scc4 cohesin loading complex⁸, from where it appears to slide to its more permanent locations. But even after sister chromatid cohesion is established, changes in transcription lead to repositioning of cohesin. Thus the sites of cohesin binding

and therefore probably sister chromatid cohesion, a key architectural feature of mitotic chromosomes, display surprising flexibility. Cohesin localization to places of convergent transcription is conserved in fission yeast, suggesting that it is a common feature of eukaryotic chromosomes.

Cohesin association with yeast and human chromosomes has been studied^{4,9–15}, but the defining characteristics of association sites, and how cohesin gets to these sites, remained unclear. We analysed cohesin binding to chromosome VI of the budding yeast *Saccharomyces cerevisiae* by chromatin immunoprecipitation (ChIP) followed by hybridization to a high-density oligonucleotide array¹⁶. The pattern of association in metaphase was similar for all cohesin subunits analysed, Scc1, Scc3, Smc3 and Pds5 (Fig. 1, and Supplementary Fig. S1). It was also similar before the establishment of sister chromatid cohesion, in cells arrested with the replication inhibitor hydroxyurea (ref. 9, and Supplementary Fig. S1). Cohesin bound 28 distinct sites, each spanning 1–4 kilobases (kb) in width. The intensity of association varied, with the strongest peaks found around the centromere, consistent with previous analyses^{9–11}. The distance between neighbouring cohesin association sites ranged from 2 to 35 kb. Almost all cohesin association sites were centred in intergenic regions where genes from opposite strands converged (Fig. 1a), as previously suggested¹⁷. Using an additional high-density array, we also mapped the association of Scc1 with chromosomes III, IV and V (Supplementary Table 1, and Supplementary Fig. S2). 91% (276 of 304) of the cohesin association sites identified lie at intergenic regions between converging genes, and of 328 convergent intergene regions, 84% were bound by cohesin.

We searched for shared motifs in the nucleotide sequences at cohesin binding sites, but failed to identify any. We therefore wondered whether the process of transcription itself was responsible for positioning cohesin towards converging intergenes. We tested this at the strong cohesin association site between *STE2*, a highly expressed gene on chromosome VI that is non-essential for vegetative growth, and *BST1*. Deletion of the *STE2* promoter largely abolished cohesin association at this site (Fig. 2a). Transcription therefore appears to direct the accumulation of cohesin downstream of the gene, and not the nucleotide sequence at this site that remained unchanged.

Deletion of the *STE2* promoter might have introduced unwanted changes to chromosomal architecture, so we addressed whether physiological changes in transcription would have similar effects on cohesin localization. The *MSH4* gene, close to the centromere of chromosome VI, is not expressed during vegetative growth, but upregulated upon entry into meiosis¹⁸. The gene is covered by cohesin in mitosis. During meiosis, cohesin over *MSH4* shifted towards the 3'-end of the gene, suggesting that *MSH4* transcription causes cohesin relocation (Fig. 2b). This further suggests that even the abundant cohesin around centromeres is positioned by transcription. We visualized meiotic cohesin by ChIP against Rec8, a meiosis-specific Scc1 homologue. Rec8 colocalized with Scc1 that also remained detectable in meiosis (Supplementary Fig. S3). Further differences compared to the mitotic cohesin pattern were all correlating with known transcriptional alterations in meiosis¹⁸. For example, YFR022w expression is downregulated in meiosis, and the locus now was covered by cohesin extending from the neighbouring *PES4/YFR024c* association site (Supplementary Fig. S3). Thus, decreased transcription allowed cohesin to occupy otherwise inaccessible space along the gene.

We also analysed *HSP30* on chromosome III, a gene strongly induced after heat-shock¹⁹. In a metaphase arrested culture at 23 °C, cohesin covered much of *HSP30*, but after shift of the culture to 37 °C for 15 min, *HSP30* was free of cohesin and an increased signal was observed downstream of *HSP30* at the neighbouring *MAK32/PET18* site (Fig. 2c, and Supplementary Fig. S4). This indicates that even after cohesin is already loaded onto chromosomes, and sister chromatid cohesion is established, transcription has the ability to

relocate cohesin. Another gene on chromosome III, *SRO9*, is downregulated in response to heat-shock¹⁹. Correspondingly, we found that cohesin covered *SRO9* after, but not before, the shift to 37 °C (Fig. 2c). These results can be explained if cohesin rings encircle and are free to slide along chromatin. The elongating transcription machinery may push cohesin towards the 3' end of transcriptional units, and recurring transcription may be required to maintain cohesin at these sites. Alternatively an event correlating with transcription, for example, a certain type of chromatin modification, may be responsible for positioning cohesin.

We next addressed where cohesin is first loaded onto chromosomes after its synthesis in late G1. Cohesin interacts with, and its loading onto chromosomes depends on, the Scc2/Scc4 complex^{8,20,21}. Cytological studies suggested that cohesin does not colocalize with Scc2/Scc4 on chromosomes, which made the molecular function of this complex difficult to rationalize. We analysed the localization of Scc2/Scc4 along chromosome VI in cells arrested in early S-phase with the replication inhibitor hydroxyurea (Fig. 3a). Scc2 and Scc4 showed an identical pattern of localization, next to both telomeres, at the centromere, as well as at numerous places along chromosome arms. Association sites differed in width and, as expected, they were distinct from the sites occupied by cohesin. A similar pattern of Scc4 localization was seen in G1 and metaphase arrested cells (Supplementary Fig. S5). The strong Scc2/Scc4 binding site at the centromere, most prominent in S-phase cells, may be responsible for loading of abundant cohesin around this locus. Comparison of Scc2/Scc4 localization to transcriptional activity along chromosome VI revealed that at some places Scc2/Scc4 binding coincided with strong transcription (Fig. 3b). Statistical tests suggest that transcriptional activity may correlate with the binding of Scc2/Scc4 (Supplementary Note 1). This describes the chromosomal distribution of the Scc2/Scc4 complex.

To investigate the role of the Scc2/Scc4 complex in cohesin loading, we analysed cohesin association in *scc2-4* mutant cells. We were surprised to find cohesin now at the sites normally occupied by the Scc2/Scc4 complex (Fig. 3c, and Supplementary Fig. S6). This suggests that an early step in the loading of cohesin onto DNA, before the step disrupted by the *scc2-4* mutation, takes place at Scc2/Scc4 binding sites. Cohesin association under these conditions may be weak, which could explain why other methods failed to reveal it⁸. We next analysed early events of cohesin loading in wild-type cells traversing synchronously through G1. Cells were arrested with the mating pheromone α -factor and released at 16 °C to slow progression through G1. The first detectable association of cohesin with chromosomes, 10 min after release, overlapped with Scc2/Scc4 binding sites (Fig. 3c, and Supplementary Fig. S6). After 30 min, cohesin became detectable also at the expected converging intergenic sites. This is consistent with the idea that cohesin is loaded onto chromosomes at Scc2/Scc4 binding sites, and thereafter relocates to more permanent places.

To see whether cohesin localization by converging transcription was a general feature of eukaryotic chromosomes, we performed ChIP of cohesin in the fission yeast *Schizosaccharomyces pombe*⁴, followed by hybridization to a high-density oligonucleotide array covering its chromosomes 2 and 3. Compared to budding yeast, genome organization and chromatin biology in fission yeast resembles more closely the situation in higher eukaryotes. We found that the fission yeast Scc1 homologue, Rad21, preferentially localized to regions of convergent transcription, even though the peaks of association often extended over larger regions (Fig. 4, and Supplementary Note 2). Therefore, cohesin distribution in response to transcription appears to be a conserved mechanism.

Sister chromatid cohesion along chromosome arms is crucial for DNA repair by homologous recombination²². The prominence of

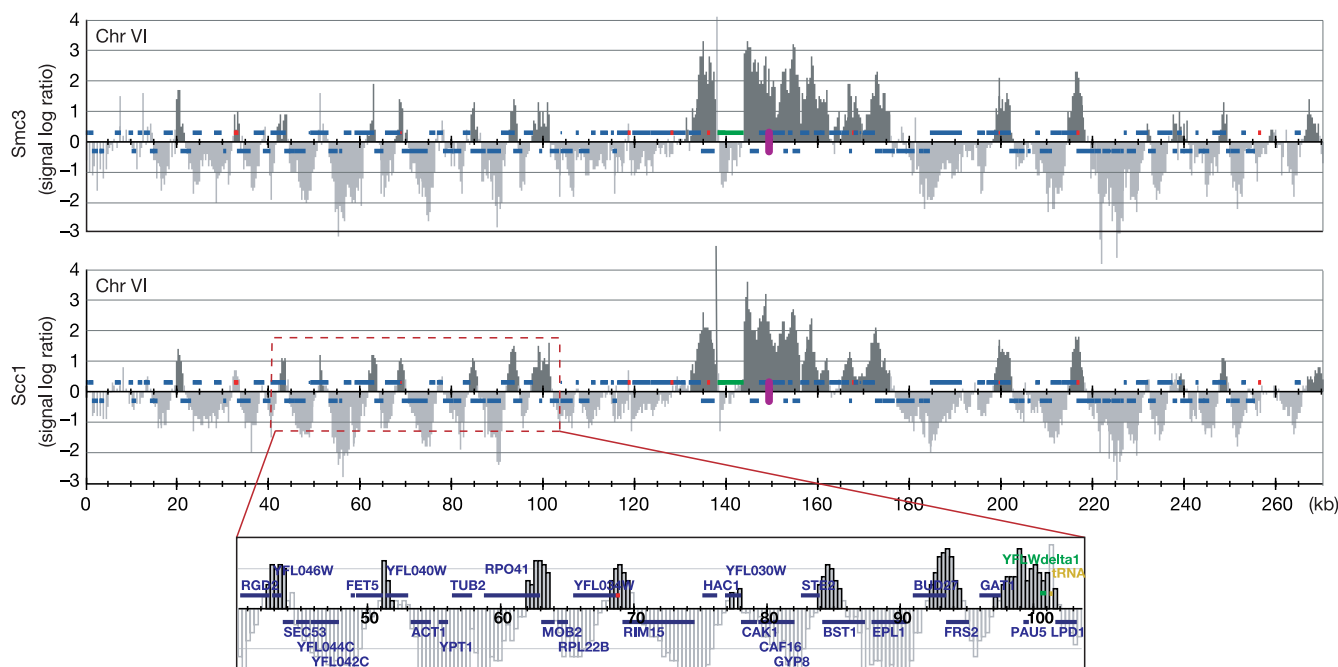


Figure 1 Cohesin localizes to convergent intergene regions along budding yeast chromosome VI. Cells containing Smc3-Flag₃ or Scc1-HA₆ were arrested in metaphase by nocodazole treatment and processed for ChIP against the epitope tagged subunits. Enrichment in the immunoprecipitated fraction relative to a whole genome DNA sample is shown along the length of the chromosome. Each bar represents the average of 16 oligonucleotide probes within adjacent 300 bp windows. The y-axis scale is log₂ Dark

grey signals represent significant binding, as detailed in the Supplementary MIAME Document. Blue bars above and below the midline are genes transcribed respectively from left to right, and from right to left. The purple oval represents the centromere. Origins of replication are depicted in red, tRNA genes in yellow, and a Ty2 transposon in green.

cohesin at converging intergenic sites therefore raises the question of whether genes are arranged in a particular order to provide for a regular pattern of cohesin association. When we analysed the succession of left- or right-facing genes along budding yeast chromosomes III–VI, we found that it seemed random. We did not find regions on these chromosomes where the likelihood of orientation of a gene following either a left- or a right-facing gene was different from equal (Supplementary Note 3). Fortuitous clusters of genes in tandem, which have been noted before²³, are expected in this random distribution. The consequent chance distribution of convergence sites agrees with the observed wide range of distances between neighbouring cohesin association sites. Thus, an important structural feature of yeast chromosomes, the sites of cohesin association, and therefore probably sister chromatid cohesion, are spaced at random. Cohesin can act as a border of transcriptional domains²⁴. This seems now unlikely to be generally the case, as neighbouring genes in budding yeast show a high probability of co-regulation, irrespective of whether they are oriented in divergent or convergent orientation²⁵.

We show here that cohesin, a critical structural component of the eukaryotic chromosome, does not have a fixed pattern of localization. The possibility that the ring-shaped cohesin complex encircles chromatin offers an intriguing explanation for this finding⁷. After DNA is transported into the cohesin ring^{21,26}, cohesin may be able to

slide along chromatin, responding to steric requirements of transcription and perhaps other forms of chromosomal metabolism. Cohesin is loaded onto chromosomes next to its loading factor Scc2/Scc4, from where it relocates to places of convergent transcription. Scc2/Scc4 has been suggested to stimulate cohesin's ATPase activity, required to open the ring for loading onto DNA. Once loaded, prevention of further Scc2/Scc4 action on cohesin would be vital to securing stability of the ring²¹. Moving cohesin away from its loading sites, promoted by strong transcriptional activity found there, would be a safe way to achieve this. But even after sister chromatid cohesion is established cohesin is not stably trapped, and changes in transcriptional activity lead to further redistribution. This suggests that cohesin provides chromosomes with an unexpectedly flexible architecture.

This scenario does not exclude the possibility that places exist where cohesin engages in a more stable contact with chromatin. For example, fission yeast cohesin's enrichment at centromeric heterochromatin is mediated by a specific interaction with the heterochromatin protein Swi6 (ref. 13). Centromeric cohesin is needed to promote bipolar spindle attachment in mitosis^{5,27}, but in contrast to budding yeast, centromeres in fission yeast and higher eukaryotes cover large regions devoid of much transcriptional activity. Cohesin cannot therefore be confined there by convergent transcription, and during the evolution of larger centromeres cohesin may have

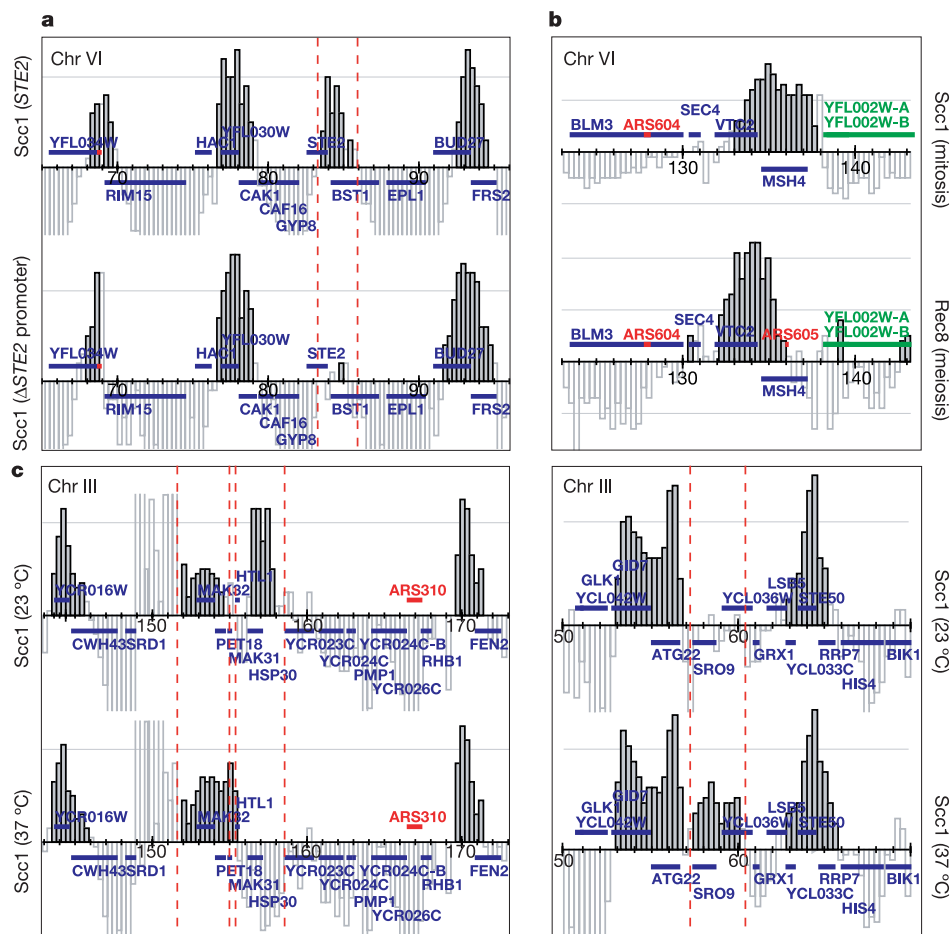


Figure 2 Cohesin is moved towards the 3' -end of genes by their transcription.

a, Localization of Scc1-HA₆ around *STE2* on chromosome VI in a nocodazole-arrested wild-type strain (top), and after deletion of the *STE2* promoter (bottom). **b**, Localization of Scc1-HA₆ around *MSH4*, close to centromere VI, in mitotic cells (top), and of Rec8-Flag₃ in meiosis (bottom) when *MSH4* is induced. **c**, Localization of Scc1-HA₆ in nocodazole-

arrested cells before (top) and 15 min after heat-shock (bottom). Regions around *HSP30* and *SRO9* on chromosome III, induced and repressed after heat-shock respectively, are shown. See Fig. 1 legend for key to colour coded chromosomal features. Red dashed lines flank the regions described in the text.

acquired an alternative way to hold on to centromeric chromatin. Stable binding to heterochromatin, versus lateral mobility, may also differentiate cohesin during unloading of a subpool in prophase when chromosomes condense and transcription ceases²⁸.

How does transcription move cohesin towards converging intergenes? Although we cannot exclude the possibility that transcriptional termination correlates with, or induces, a specific chromatin state that attracts cohesin, it is unclear how such a feature would be

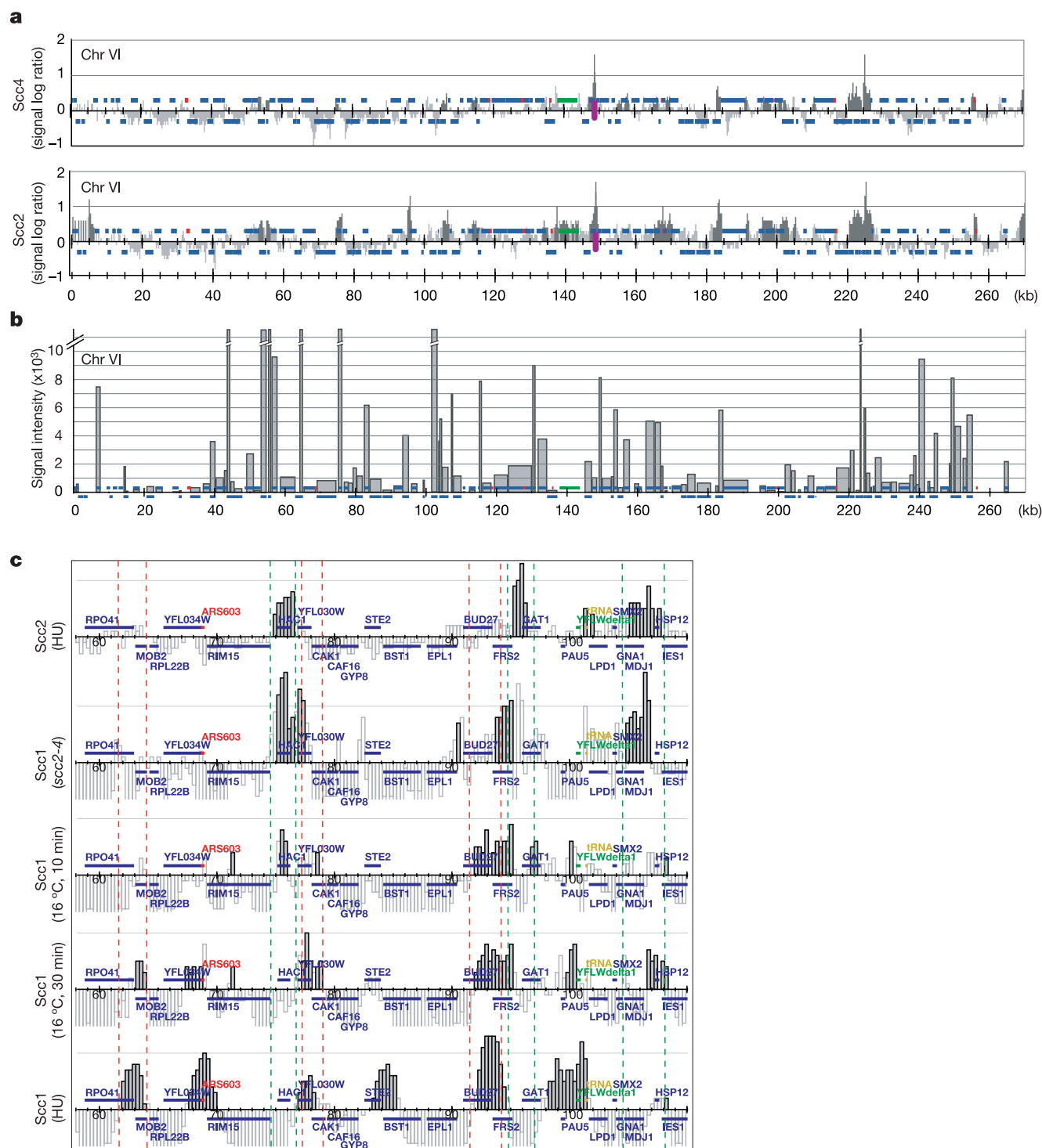


Figure 3 Cohesin loading at, and movement away from, sites of Scc2/Scc4 binding. **a**, Localization of Scc4-HA₆ and Scc2-HA₆ along chromosome VI in hydroxyurea-arrested cells. See Fig. 1 legend for meaning of colour coded chromosomal features. **b**, Transcriptional activity along chromosome VI in wild-type logarithmically growing cells. The signal intensity reflects the abundance of transcripts, as an approximation of transcriptional activity. Broken bars exceed 10,000; see Supplementary Fig. S5 for full values. **c**, Detail of the localization on the left arm of chromosome VI, from top to bottom, of

Scc2-HA₆ in hydroxyurea-arrested (HU) cells, Scc1-HA₆ in *scc2-4* mutant cells released from G1 into hydroxyurea block at restrictive temperature, Scc1-HA₆ in wild-type cells 10 min, and 30 min following release from G1 at 16 °C, and Scc1-HA₆ in hydroxyurea-arrested wild-type cells. The green and red dashed lines flank sites usually seen bound by Scc2 and Scc1, respectively. See Fig. 1 legend for meaning of colour coded chromosomal features.



577

17. Filipiński, J. & Mucha, M. Structure, function and DNA composition of *Saccharomyces cerevisiae* chromatin loops. *Gene* **300**, 63–68 (2002).
18. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
19. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
20. Tóth, A. *et al.* Yeast cohesin complex requires a conserved protein, Eco1p (Ctf7), to establish cohesion between sister chromatids during DNA replication. *Genes Dev.* **13**, 320–333 (1999).
21. Arumugam, P. *et al.* ATP hydrolysis is required for cohesin's association with chromosomes. *Curr. Biol.* **13**, 1941–1953 (2003).
22. Sjögren, C. & Nasmyth, K. Sister chromatid cohesion is required for postreplicative double-strand break repair in *Saccharomyces cerevisiae*. *Curr. Biol.* **11**, 991–995 (2001).
23. Jacq, C. *et al.* The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV. *Nature* **387** (suppl), 75–78 (1997).
24. Donze, D., Adams, C. R., Rine, J. & Kamakaka, R. T. The boundaries of the silenced *HMR* domain in *Saccharomyces cerevisiae*. *Genes Dev.* **13**, 698–708 (1999).
25. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* **26**, 183–186 (2000).
26. Weitzer, S., Lehane, C. & Uhlmann, F. A model for ATP hydrolysis-dependent binding of cohesin to DNA. *Curr. Biol.* **13**, 1930–1940 (2003).
27. Toyoda, Y. *et al.* Requirement of chromatid cohesion proteins Rad21/Scc1 and Mis4/Scc2 for normal spindle-kinetochore interaction in fission yeast. *Curr. Biol.* **12**, 347–358 (2002).
28. Waizenegger, I. C., Hauf, S., Meinke, A. & Peters, J.-M. Two distinct pathways remove mammalian cohesin complexes from chromosome arms in prophase and from centromeres in anaphase. *Cell* **103**, 399–410 (2000).
29. Reid, R. J. D., Sunjevaric, I., Kedache, M. & Rothstein, R. Efficient PCR-based gene disruption in *Saccharomyces* strains using intergenic primers. *Yeast* **19**, 319–328 (2002).
30. Ohta, K. *et al.* Mutations in the MRE11, RAD50, XRS2, and MRE2 genes alter chromatin configuration at meiotic DNA double-stranded break sites in premeiotic and meiotic cells. *Proc. Natl Acad. Sci. USA* **95**, 645–651 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are indebted to E. Schwob for initiating this collaboration. We also thank A. Nakada and T. Chaplin for technical support, R. Rothstein for reagents, and J. Cau, J. Sgouros, J. Svejstrup and members of our laboratories for discussions and comments on the manuscript. A.L. was supported by an EU Marie Curie individual fellowship and a *Journal of Cell Science* travelling fellowship; K.S. acknowledges support through a MEXT grants-in-aid for priority areas; E.U. was supported by the EMBO Young Investigator programme.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.U. (frank.uhlmann@cancer.org.uk).

Structural determinants for generating centromeric chromatin

Ben E. Black^{1,2}, Daniel R. Foltz^{1,2}, Srinivas Chakravarthy⁴, Karolin Luger⁴, Virgil L. Woods Jr³ & Don W. Cleveland^{1,2,3}

¹Ludwig Institute for Cancer Research, ²Department of Cellular and Molecular Medicine, and ³Department of Medicine, University of California, San Diego, La Jolla, California 92093, USA

⁴Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, Colorado 80523, USA

Mammalian centromeres are not defined by a consensus DNA sequence. In all eukaryotes a hallmark of functional centromeres—both normal ones and those formed aberrantly at atypical loci—is the accumulation of centromere protein A (CENP-A), a histone variant that replaces H3 in centromeric nucleosomes^{1–7}. Here we show using deuterium exchange/mass spectrometry coupled with hydrodynamic measures that CENP-A and histone H4 form sub-nucleosomal tetramers that are more compact and conformationally more rigid than the corresponding tetramers of histones H3 and H4. Substitution into histone H3 of the domain of CENP-A responsible for compaction is sufficient to direct it to centromeres. Thus, the centromere-targeting domain of CENP-A confers a unique structural rigidity to the nucleo-

somes into which it assembles, and is likely to have a role in maintaining centromere identity.

Centromeres and their associated kinetochores are protein–DNA complexes that mediate spindle microtubule attachment during mitosis and are necessary to mediate delivery of one copy of each chromosome to each daughter cell (for review see ref. 1). Mammalian centromeres contain megabases of repetitive, α -satellite DNA that is organized into specialized chromatin consisting of nucleosomes in which histone H3 is replaced by the variant CENP-A. Removal of CENP-A reduces fidelity of chromosome segregation^{8–13}, leading to the proposal that centromere-specific, CENP-A-containing nucleosomes may recruit other centromere and kinetochore components, at least within the context of an array of CENP-A-containing nucleosomes that generates a higher-order chromatin structure required for the formation of the inner kinetochore surface¹⁴. CENP-A represents the most likely candidate to be the principal epigenetic ‘mark’ that maintains the identity of centromeres after DNA replication in organisms containing complex, multisubunit centromeres, such as humans^{5,15}. Maintaining centromere identity during normal cell division may be its most important role, but how it does this and even the process by which CENP-A is targeted to centromeres remains unclear.

Although CENP-A was originally identified nearly 20 yr ago^{16,17}, no evidence has yet emerged on how its presence in nucleosomes affects the properties of those nucleosomes. We have now tested this by co-purification of a complex of recombinant CENP-A and histone H4 and comparison of the complex to the well-defined H3–H4 heterotetramer (Fig. 1a). As with H3–H4, the CENP-A–H4 complex has apparent equal stoichiometry of its subunits. Size exclusion chromatography and analytical ultracentrifugation were used to determine the composition of the native CENP-A–H4 complex (Fig. 1b, c). Although the CENP-A polypeptide is four amino acids larger than H3 (140 versus 136 amino acids), the CENP-A–H4 complex chromatographs as a single species with a Stokes’ radius of 28 Å (Fig. 1b), a smaller value than for H3–H4 (30.5 Å). This is consistent with CENP-A–H4 existing as a compact heterotetramer or an extremely irregular heterodimer. Analytical ultracentrifugation (Fig. 1c) demonstrated that the CENP-A–H4 complex sedimented as a single species at 3.2S (where S is a sedimentation coefficient), as did H3–H4. Taken together, these data indicate a native complex of 52 kDa (see Methods); that is, a tetramer of two copies each of CENP-A and H4 that is more compact than the corresponding H3–H4 heterotetramer (59 kDa). In contrast, H2A–H2B, with respective amino acid contents of 130 and 126 amino acids, exists as a dimer of 25 kDa (20.5 Å Stokes’ radius and 2.1 S sedimentation coefficient).

Deuterium exchange/mass spectrometry (DXMS; Fig. 1d, e) was used to determine whether the apparently more compact CENP-A–H4 tetramer is a conformationally more rigid structure than histone H3–H4. Under native conditions, deuterium was exchanged from heavy water to accessible protons within either tetrameric complex. This technique measures only the peptide amide proton exchange, so an increase in deuteration directly corresponds to exchange along the backbone of the polypeptide¹⁸, allowing measurement of the solvent accessibility at all regions within a polypeptide chain¹⁹. The complexes were then denatured, proteolysed, and the deuterium content of each peptide determined by mass spectrometry. For both H3–H4 and CENP-A–H4 complexes, a wide range of exchange rates was identified in peptides that covered 97% and 98% of the polypeptide backbone, respectively. The overall deuterium exchange profiles of CENP-A and H3 were very similar, with the single exception that much slower exchange was observed in residues 94–116 of CENP-A compared with the corresponding region in H3 (Fig. 1d, e). Additionally, CENP-A induced conformational alterations in H4, as revealed by two separate regions (residues 58–78 and 86–91) that were of much lower (~10-fold) accessibility to exchange when complexed with CENP-A as com-

pared with H3. One of these regions in H4 (residues 58–78) remained rigidly packed when in complex with CENP-A, with essentially no exchange in the amide hydrogens during 100 s, and only a small proportion exchanged by 1,000 s (Fig. 2a). However, the same domain of H4 in a tetramer complex with H3 fully exchanged deuterium by 1,000 s. A small proportion of this region in H3–H4 tetramers that exchanged at the earliest time points (100 s) did so essentially completely, yielding a bi-modal distribution of peptides with unexchanged or fully exchanged hydrogens, a behaviour also seen in 13 other overlapping peptides from this region. This indicates a rapid, cooperative ‘all or none’ unfolding between open and closed conformations in this domain in the H3–H4

tetramers, at times when CENP-A–H4 remains rigidly packed with minimal exchange.

The deuterium exchange profiles were positioned onto the existing crystallographic model of H3–H4 (isolated from the nucleosomal structure²⁰) and onto a corresponding model of CENP-A–H4 based on this structure (Fig. 2b, c). The interaction domain between H3 and H4 includes the $\alpha 2$ helix from H3 and the $\alpha 2$ and $\alpha 3$ helices from H4. The region surrounding the H4 interface with H3, consisting of a portion of the $\alpha 2$ helix and the L1 linker, is highly conserved in H3 proteins from diverse species (H3 of human and frog is 95% identical throughout L1 and the $\alpha 2$ helix) with no variation in spacing of residues (Fig. 2d). CENP-A

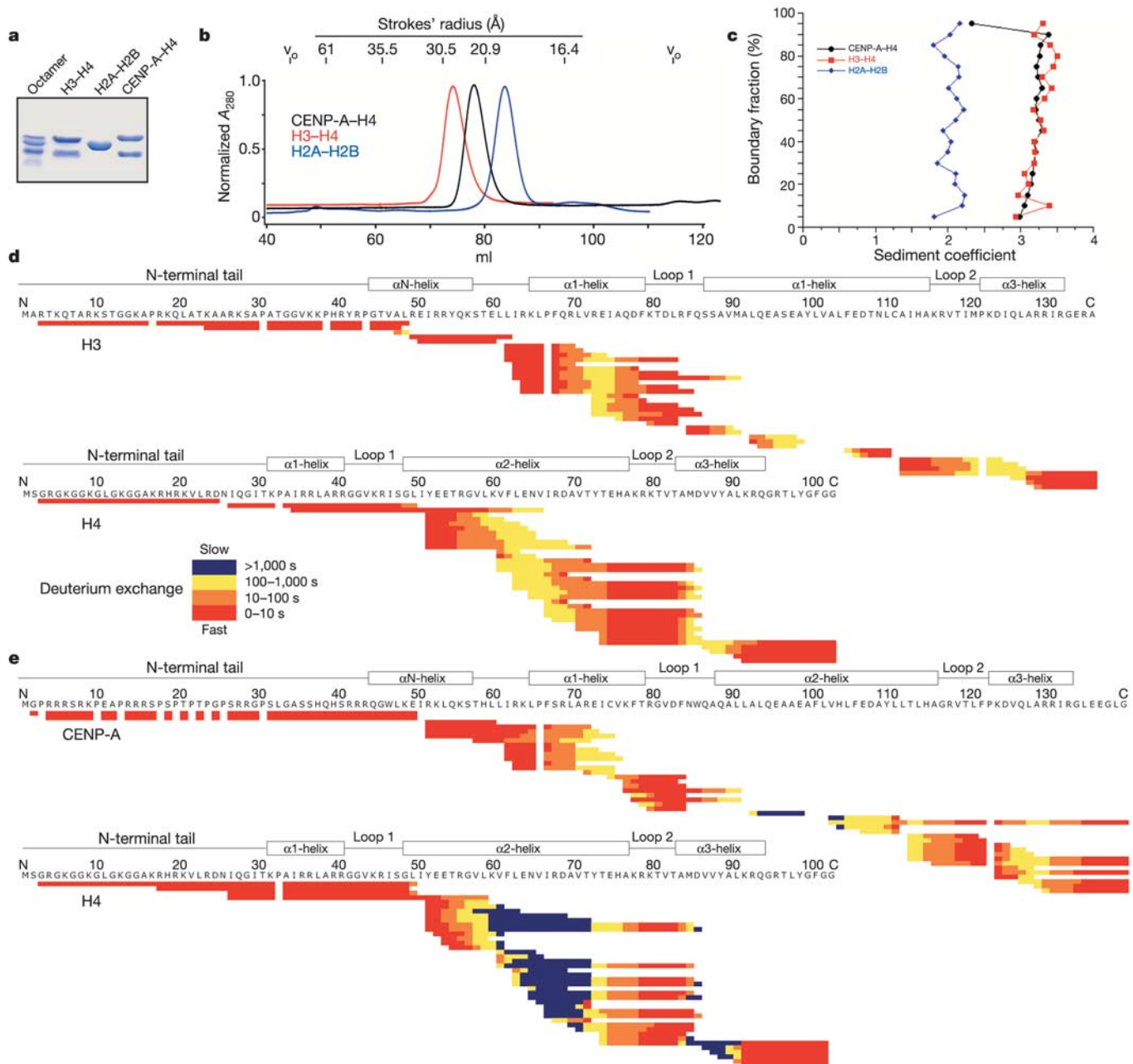


Figure 1 The CENP-A–H4 complex is a compact heterotetramer. **a**, Purified histone complexes. **b**, Purified complexes and standards of known Stokes' radii were analysed by Superdex 200 chromatography under identical conditions. The x axis indicates the elution volume in millilitres (ml). **c**, Analysis of histone complexes by analytical

ultracentrifugation. **d**, **e**, DXMS analysis of the H3–H4 (**d**) and CENP-A–H4 (**e**) heterotetramers. Known (**d**) or predicted (**e**) secondary structural features are indicated above the primary sequence.

proteins, however, vary greatly from H3 in these regions: human H3 and CENP-A are only 51% identical, and the spacing is altered by inserted residues in L1 (Fig. 2d). This region is also the site of mutation in several temperature-sensitive alleles of *CSE4*, the budding yeast orthologue of CENP-A, that were identified in a mutagenesis-based screen for viability²¹. Moreover, the sites where CENP-A varies from H3 are well conserved among CENP-A orthologues from diverse species, all of which suggests that the unique H4 interface on human CENP-A (Fig. 2c, d) has been maintained throughout centromeric chromatin evolution. Consistent with this, human CENP-A containing a substituted L1 linker (CENP-A(L1); Supplementary Fig. S1) and the amino-terminal and central part of the $\alpha 2$ helix (CENP-A(H2.1) and CENP-A(H2.2), respectively; Supplementary Fig. S1) of H3 (ref. 22) did not associate with centromeres, and instead incorporated all along chromosome arms in detergent-resistant chromatin in interphase nuclei and in mitotic chromosomes.

The behaviour of these CENP-A mutant proteins suggested that L1 and the $\alpha 2$ helix specify CENP-A targeting to the centromere. To test this hypothesis, L1 and the $\alpha 2$ helix of H3 were replaced by the corresponding regions from CENP-A (Fig. 3a, green). This produced a chimaeric H3 variant with 22-amino-acid substitutions from CENP-A that targeted to centromeres in interphase (Fig. 3b) and throughout mitosis (Fig. 3c). Determination (from 40 to 50

cells) of the centromere/nucleoplasmic ratios of each of these yellow fluorescent protein (YFP) fusions showed that the H3/CENP-A chimera was targeted to centromeres to a similar degree as CENP-A ($85 \pm 19\%$ ($\pm$ s.e.m.) of the selectivity of wild-type CENP-A). This is in contrast with H3, which was de-enriched by $7 \pm 2\%$ at centromeres. Thus, L1 and $\alpha 2$, neither of which alone are able to target H3 to centromeres^{22,23}, together form the CENP-A targeting domain (CATD), which is sufficient to restrict H3 targeting to centromeres. Combined with previous findings that mutation in every other region of CENP-A can be carried out without abolishing centromeric targeting²², the finding that this region confers centromeric localization to H3 (Fig. 3b, c) strongly indicates that the CATD contains the *cis*-acting information required for proper deposition at centromeres. Both parts of the CATD work together to position the interaction surface that restricts the conformational flexibility of H4. This seems to be a general model for targeting histone H3 variants. A three-amino-acid divergence within the $\alpha 2$ helix of the histone H3 variant H3.3 (Fig. 2d, blue lettering) confers replication-independent deposition²⁴, which, as with CENP-A, is independent of DNA replication and exhibits a discrete sub-chromosomal localization pattern^{24,25}.

Recombinant H3(CATD) formed pre-nucleosomal complexes in a 1:1 stoichiometry with H4 (Fig. 3d–g), just as previously seen for H3 and CENP-A (Fig. 1a). DXMS of the H3(CATD)–H4 complex

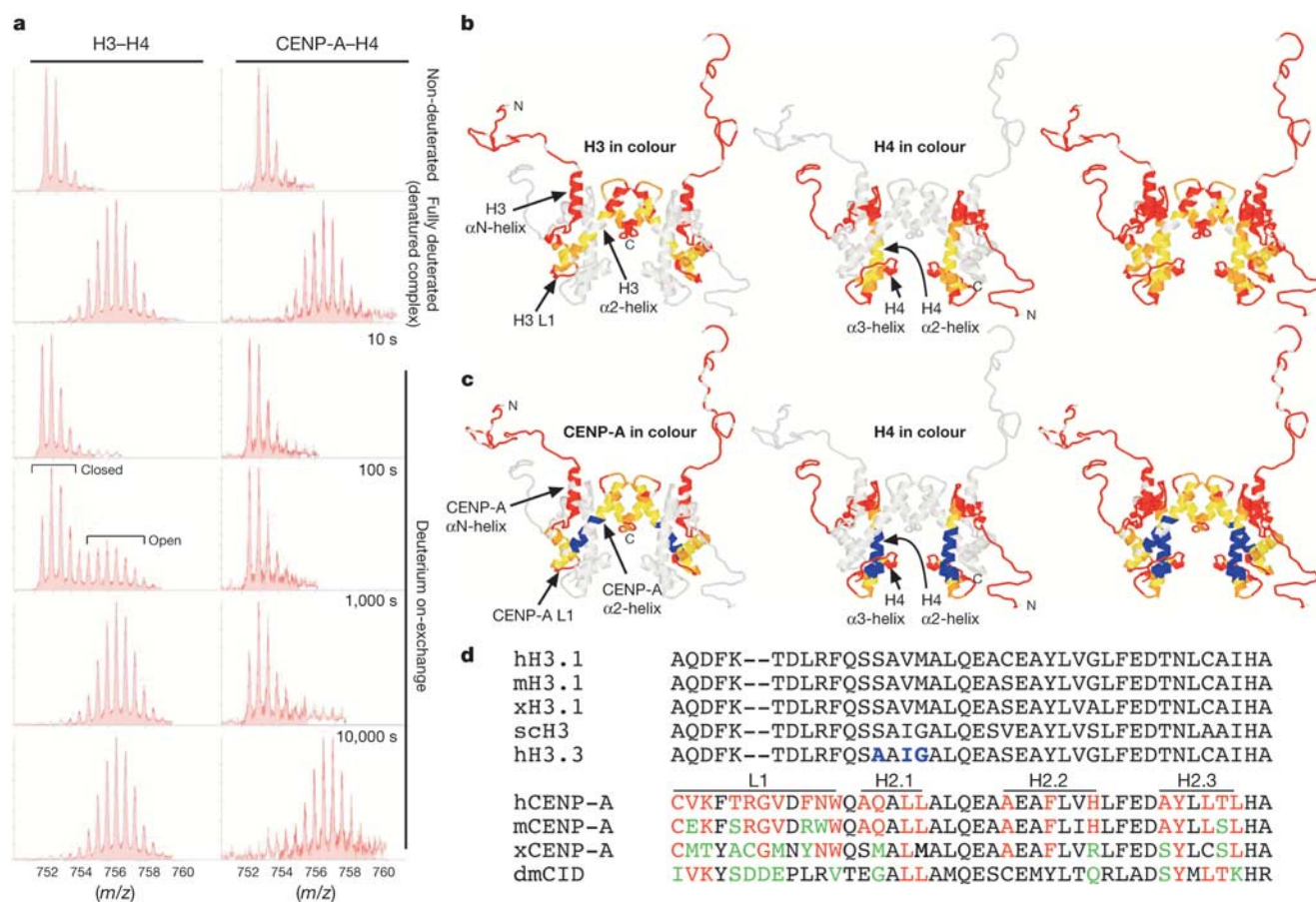


Figure 2 Unique structural features at the CENP-A–H4 interface. **a**, A peptide (in the $[+2]$ charge state), corresponding to a portion of the $\alpha 2$ helix from histone H4 (residues 60–72), displays delayed deuteration exchange when bound to CENP-A versus histone H3. The x axes are plotted as mass:charge (m/z). **b**, **c**, The consensus deuteration exchange data were labelled at each residue of the known H3–H4 structure (**b**; Protein Data Bank number 1kx5 (ref. 20)) or a model of CENP-A–H4 (**c**; generated in SWISS-MODEL) using

the same colour scheme as in Fig. 1. **d**, Alignment of L1 and $\alpha 2$ helix from histone H3 and H3 variants. dm, *Drosophila melanogaster*; h, human; m, mouse; sc, *Saccharomyces cerevisiae*; x, *Xenopus*. Deviation from histone H3 in human CENP-A (red lettering) is conserved at many sites in CENP-A orthologues, or there is another substitution at these positions (green lettering) that also differs from H3 sequences.

revealed regions of slow deuterium exchange on both H3(CATD) and H4 (Fig. 3e) that localized to the interface between H3 and H4 in the nucleosomal model (compare Fig. 2c with 3f). Furthermore, like CENP-A itself, H3(CATD) conferred a markedly reduced conformational flexibility to the domain of H4 in direct contact with it relative to the same H4 domain when in complex with wild-type H3 (Fig. 3g). Thus, the CATD not only confers centromere targeting to H3, but before deposition into chromatin it is also sufficient to mediate reduction in conformational flexibility (compaction) in both H3 and H4, assembled in the absence of CENP-A, in a pre-nucleosomal H3–H4 complex.

The presence of CENP-A has been used to define functional centromeres from all species, including those centromeres formed *de novo* on long arrays of α -satellite DNA^{2–4} as well as normally non-centromeric loci that acquire centromere function after aberrant chromosomal rearrangements^{5–7}. These latter ‘neocentromeres’ form at loci with no recognizable centromeric or repetitive DNA sequences. This has led to the proposal that CENP-A serves as an epigenetic mark for propagation of centromere identity. The initial events that lead to neocentromere formation are, however, unknown. Our effort here has now provided a structural and

biochemical basis for a structurally and functionally distinguishing epigenetic mark that propagates centromere identity independent of DNA sequence (see model in Supplementary Fig. S2). Although the continued presence of endogenous CENP-A does not allow us to exclude completely involvement of other regions of CENP-A in centromere targeting or templating, substitution into H3 of the CENP-A CATD domain generates both selective centromere targeting *in vivo* and a compact, rigid conformation in the absence of CENP-A *in vitro*. Because CENP-A is necessary for kinetochore assembly at mitosis^{8–13}, it seems likely that this unique chromatin environment contributes to centromere recognition by downstream kinetochore components. After initial deposition, the conformational rigidity of CENP-A nucleosomes not only serves as a template to restrict assembly of new CENP-A to existing chromosomal loci, but also confers a more rigid platform on which other components of functional centromeres can be deposited. We propose that CENP-A is deposited on centromeric DNA by one or more unidentified chromatin assembly factors²⁶, where specificity is conferred by the pre-established CENP-A centromere mark and the *cis*-acting CATD. As long as some remains during DNA replication, either by incomplete removal or by only transient

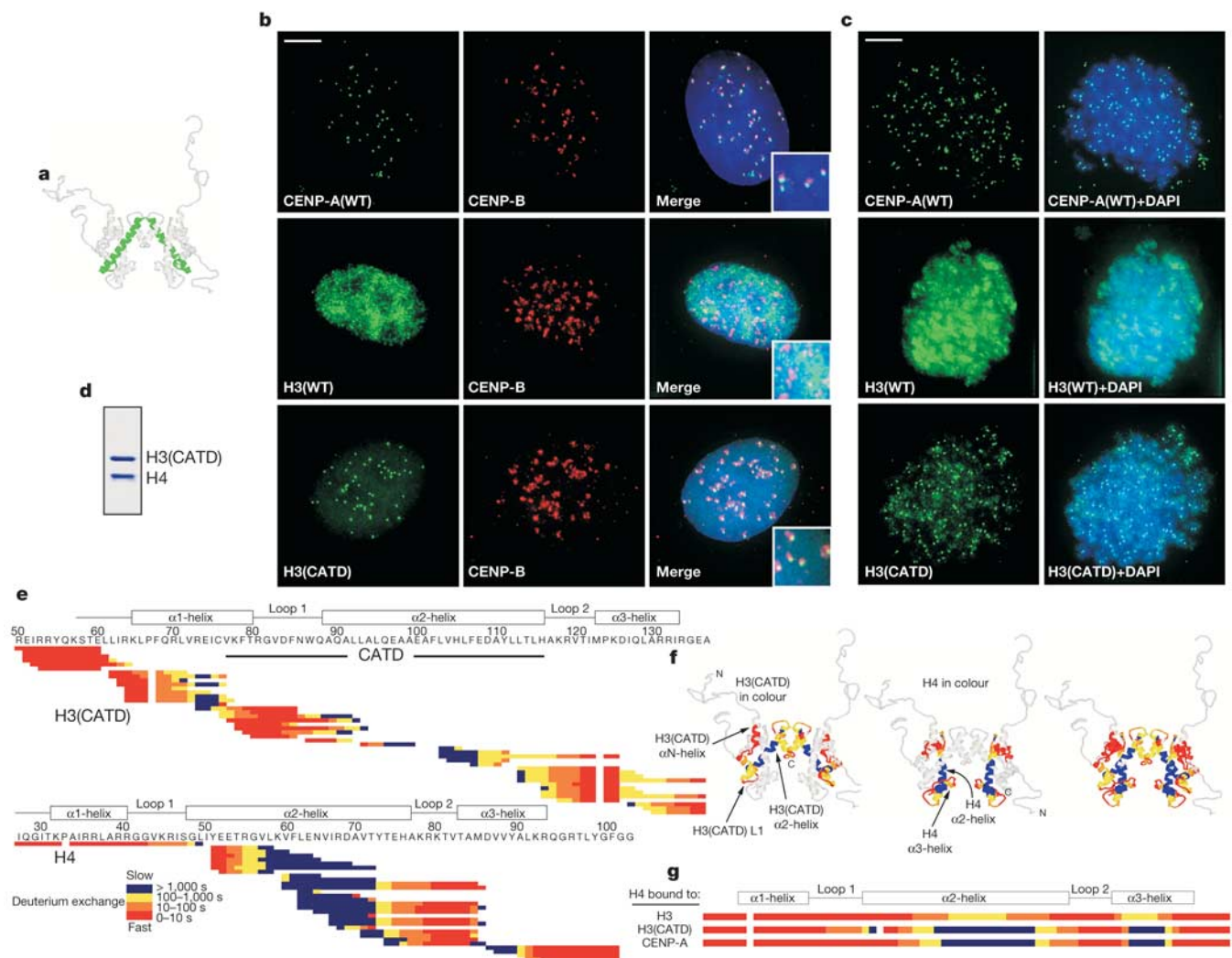


Figure 3 Converting H3 to a centromeric histone. **a**, The chimaeric H3 protein containing the CATD (L1 and the α 2 helix; green) is shown in the context of the H3–H4 structural model²⁰. **b**, Targeting of H3(CATD) to the centromere in interphase cells. DNA was stained with DAPI (blue). **c**, Targeting of H3(CATD) to mitotic kinetochores. Scale bar, 5 μ m.

d, Purified H3(CATD)–H4 complex. **e**, DXMS analysis of the H3(CATD)–H4 complex. **f**, Mapping of the consensus DXMS profile of H3(CATD)–H4. **g**, Alignment of the consensus DXMS profiles for histone H4 when complexed with H3, H3(CATD), or CENP-A.

release during replication, the specialized chromatin environment with compact, less flexible CENP-A nucleosomes both comprises and perpetuates the epigenetic mark.

Our results support a model of establishment and maintenance of human centromeres dictated by unique structural features within the CENP-A–H4 interface that direct the first step in centromere formation. In such a model, the rare formation of new centromeres that accompanies chromosomal rearrangements may arise by acquisition of a small cluster of CENP-A by a non-centromeric locus of a chromosomal fragment. This alters local chromatin structure, and the nascent centromere is extended and maintained by a self-propagating mechanism for CENP-A deposition. □

Methods

Purification and biochemical analysis of recombinant histones

A bacterial codon-biased synthetic human CENP-A gene was constructed using polymerase chain reaction (PCR) from overlapping oligonucleotides. This was co-expressed with synthetic histone H4 (ref. 27) from a bi-cistronic vector derived from pET-12 (Novagen) in BL21 [DE3] (pLysS). The synthetic H4 gene product is identical in sequence to human H4. The soluble CENP-A–H4 complex was purified to near homogeneity by sequential hydroxyapatite and SP (cation exchange) chromatography. Purification of recombinant H2A–H2B heterodimers and H3–H4 heterotetramers was performed as described²⁸. Gel filtration was performed on a Superdex 200 column that was calibrated with standards of the indicated Stokes' radii (Amersham). Maximal absorbance readings at 280 nm (A_{280}) were normalized to one, and peak fractions were analysed by SDS–polyacrylamide gel electrophoresis and Coomassie blue staining to confirm that the complexes remained intact (data not shown). Analytical ultracentrifugation was performed in a Beckman Optima XL-1 using an AN60 rotor at 50,000 r.p.m. Raw data were analysed with Ultrascan software, and the experimentally derived molecular weights for histone complexes were calculated using partial-specific volumes of 0.710–0.753 cm³ g^{−1}, which covers the normal range of partial-specific volume for most proteins and the higher values reported by others for histone complexes²⁹. Values are plotted as sedimentation coefficient versus the per cent boundary fraction. The molecular weights reported here were calculated using a partial-specific volume of 0.710 for all histone complexes.

DXMS

For a detailed description of protein fragmentation mapping, deuterium exchange and analysis of exchange profiles, see Supplementary Information.

Mammalian expression vectors

Constructs expressing CENP-A (wild type and mutants) tagged with a triple haemagglutinin (HA) epitope²² were provided by K. Sullivan, with the exception of CENP-A(H2.2)–HA. This mutation was generated by mutagenic PCR using the Quickchange system (Stratagene) with the CENP-A(WT) HA construct as the template. YFP fusion proteins for CENP-A(WT) and H3(WT) were constructed in a derivative of pEYFP (Clontech) where constitutive expression of reporter constructs are driven from the CMV promoter. YFP–H3(CATD) was constructed by generating an upstream EcoRV site and a downstream MluI site in pYFP–H3(WT) by sequential rounds of Quickchange mutagenesis. Digestion of this intermediate construct with EcoRV and MluI excised the coding region corresponding to L1 and the $\alpha 2$ helix of H3. This was replaced with a PCR fragment, with engineered PvuII (5') and MluI (3') sites, that contained L1 and the $\alpha 2$ helix of CENP-A. The resulting construct was sequenced to confirm the coding sequence of H3(CATD).

Immunofluorescence

T98G cells (human glioblastoma) were grown on glass coverslips in DMEM containing 10% fetal bovine serum and transfected with the indicated plasmids using Effectene (Qiagen). For analysis of interphase centromeres, coverslips were processed for indirect immunofluorescence two days after transfection. For mitotic chromosome preparations, cells were grown two days post-transfection and then colcemid (250 ng ml^{−1}; Gibco) was added, and cells were incubated for an additional 4–5 h. Mitotic cells were collected by agitation, hypotonically swollen in 75 mM KCl, spun onto glass coverslips using a Sorvall HS4 rotor at 6,000 r.p.m. for 30 min at room temperature, and then processed for indirect immunofluorescence. Either unlabelled or fluorescein-isothiocyanate-labelled anti-HA mouse monoclonal antibodies (HA.11; Covance) were used at 1 μ g ml^{−1}. The anti-CENP-B mouse monoclonal antibody was prepared as tissue culture supernatant from 2D-7 hybridomas³⁰ obtained from the American Type Culture Collection and used at 1:4. Donkey anti-mouse Cy5-conjugated and goat anti-mouse fluorescein-isothiocyanate-conjugated secondary antibodies were obtained from Jackson ImmunoResearch Laboratories. Samples were stained with 4,6-diamidino-2-phenylindole (DAPI) before mounting with Vectashield medium (Vector Laboratories). Digital images were captured using Deltavision Softworx software by a Roper Coolsnap interline charge-coupled device camera mounted on an Olympus IX-70 inverted microscope. For each sample, images were collected at 0.2 μ m z-sections that were subsequently deconvolved using identical parameters. The z-stacks were then projected as single two-dimensional images and assembled using Adobe Photoshop (version 7.0) and Macromedia Freehand

(version 10.0). All images shown are representative of results obtained in multiple experiments.

Received 4 March; accepted 15 June 2004; doi:10.1038/nature02766.

- Cleveland, D. W., Mao, Y. & Sullivan, K. F. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* **112**, 407–421 (2003).
- Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. Genomic and genetic definition of a functional human centromere. *Science* **294**, 109–115 (2001).
- Ohzeki, J., Nakano, M., Okada, T. & Masumoto, H. CENP-B box is required for *de novo* centromere chromatin assembly on human alphoid DNA. *J. Cell Biol.* **159**, 765–775 (2002).
- Rudd, M. K., Mays, R. W., Schwartz, S. & Willard, H. F. Human artificial chromosomes with alpha satellite-based *de novo* centromeres show increased frequency of nondisjunction and anaphase lag. *Mol. Cell Biol.* **23**, 7689–7697 (2003).
- Sullivan, B. A., Blower, M. D. & Karpen, G. H. Determining centromere identity: cyclical stories and forking paths. *Nature Rev. Genet.* **2**, 584–596 (2001).
- Depinet, T. W. *et al.* Characterization of neo-centromeres in marker chromosomes lacking detectable alpha-satellite DNA. *Hum. Mol. Genet.* **6**, 1195–1204 (1997).
- Du Sart, D. *et al.* A functional neo-centromere formed through activation of a latent human centromere and consisting of non-alpha-satellite DNA. *Nature Genet.* **16**, 144–153 (1997).
- Stoler, S., Keith, K. C., Curnick, K. E. & Fitzgerald-Hayes, M. A mutation in CSE4, an essential gene encoding a novel chromatin-associated protein in yeast, causes chromosome nondisjunction and cell cycle arrest at mitosis. *Genes Dev.* **9**, 573–586 (1995).
- Buchwitz, B. J., Ahmad, K., Moore, L. L., Roth, M. B. & Henikoff, S. A histone-H3-like protein in *C. elegans*. *Nature* **401**, 547–548 (1999).
- Howman, E. V. *et al.* Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. *Proc. Natl Acad. Sci. USA* **97**, 1148–1153 (2000).
- Takahashi, K., Chen, E. S. & Yanagida, M. Requirement of Mis6 centromere connector for localizing a CENP-A-like protein in fission yeast. *Science* **288**, 2215–2219 (2000).
- Blower, M. D. & Karpen, G. H. The role of *Drosophila* CID in kinetochore formation, cell-cycle progression and heterochromatin interactions. *Nature* **391**, 200–204 (1998).
- Goshima, G., Kiyomitsu, T., Yoda, K. & Yanagida, M. Human centromere chromatin protein hMis12, essential for equal segregation, is independent of CENP-A loading pathway. *J. Cell Biol.* **160**, 25–39 (2003).
- Blower, M. D., Sullivan, B. A. & Karpen, G. H. Conserved organization of centromeric chromatin in flies and humans. *Dev. Cell* **2**, 319–330 (2002).
- Choo, K. H. Domain organization at the centromere and neocentromere. *Dev. Cell* **1**, 165–177 (2001).
- Earnshaw, W. C. & Rothfield, N. Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma* **91**, 313–321 (1985).
- Palmer, D. K. & Margolis, R. L. Kinetochore components recognized by human autoantibodies are present on mononucleosomes. *Mol. Cell Biol.* **5**, 173–186 (1985).
- Hoofnagle, A. N., Resing, K. A. & Ahn, N. G. Protein analysis by hydrogen exchange mass spectrometry. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 1–25 (2003).
- Hamuro, Y. *et al.* Dynamics of cAPK type II β activation revealed by enhanced amide H/2H exchange mass spectrometry (DXMS). *J. Mol. Biol.* **327**, 1065–1076 (2003).
- Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
- Glowczewski, L., Yang, P., Kalashnikova, T., Santisteban, M. S. & Smith, M. M. Histone-histone interactions and centromere function. *Mol. Cell Biol.* **20**, 5700–5711 (2000).
- Shelby, R. D., Vafa, O. & Sullivan, K. F. Assembly of CENP-A into centromeric chromatin requires a cooperative array of nucleosomal DNA contact sites. *J. Cell Biol.* **136**, 501–513 (1997).
- Vermaak, D., Hayden, H. S. & Henikoff, S. Centromere targeting element within the histone fold domain of Cid. *Mol. Cell Biol.* **22**, 7553–7561 (2002).
- Ahmad, K. & Henikoff, S. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol. Cell* **9**, 1191–1200 (2002).
- Shelby, R. D., Monier, K. & Sullivan, K. F. Chromatin assembly at kinetochores is uncoupled from DNA replication. *J. Cell Biol.* **151**, 1113–1118 (2000).
- Haushalter, K. A. & Kadonaga, J. T. Chromatin assembly by DNA-translocating motors. *Nature Rev. Mol. Cell Biol.* **4**, 613–620 (2003).
- Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
- Luger, K., Rechsteiner, T. J. & Richmond, T. J. Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol.* **304**, 3–19 (1999).
- Eickbush, T. H. & Moudrianakis, E. N. The histone core complex: an octamer assembled by two sets of protein-protein interactions. *Biochemistry* **17**, 4955–4964 (1978).
- Earnshaw, W. C. *et al.* Molecular cloning of cDNA for CENP-B, the major human centromere autoantigen. *J. Cell Biol.* **104**, 817–829 (1987).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Kahana, J. Shah and K. Sullivan for reagents, and I. Cheeseman, B. Cottrell, P. Dyer, C. Gessner, F. Gordon, J. Kim, S. McBryant, D. Pantazatos and S. W. Englander for advice and technical assistance over the course of this study. We also thank J. Shah for comments on the manuscript. This research was supported by grants from the NIH to D.W.C. and V.L.W., BioStar and Life Sciences Informatics grants from the University of California and ExSar Corporation to V.L.W., and by postdoctoral fellowships from the American Cancer Society (B.E.B.) and the NIH (D.R.E.). Salary support for D.W.C. was provided by the Ludwig Institute for Cancer Research.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to D.W.C. (dcleveland@ucsd.edu) or V.L.W. (vwoods@ucsd.edu).

Structural basis for vinculin activation at sites of cell adhesion

Constantina Bakolitsa¹, Daniel M. Cohen², Laurie A. Bankston¹,
Andrey A. Bobkov¹, Gregory W. Cadwell¹, Lisa Jennings³,
David R. Critchley³, Susan W. Craig² & Robert C. Liddington¹

¹Program on Cell Adhesion, The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037, USA

²Department of Biological Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

³Department of Biochemistry, University of Leicester, Leicester LE1 7RH, UK

Vinculin is a highly conserved intracellular protein with a crucial role in the maintenance and regulation of cell adhesion and migration^{1–3}. In the cytosol, vinculin adopts a default autoinhibited conformation^{4,5}. On recruitment to cell–cell and cell–matrix adherens-type junctions, vinculin becomes activated and mediates various protein–protein interactions that regulate the links between F-actin and the cadherin and integrin families of cell-adhesion molecules. Here we describe the crystal structure of the full-length vinculin molecule (1,066 amino acids), which shows a five-domain autoinhibited conformation in which the carboxy-terminal tail domain is held pincer-like by the vinculin head, and ligand binding is regulated both sterically and allosterically. We show that conformational changes in the head, tail and proline-rich domains are linked structurally and thermodynamically, and propose a combinatorial pathway to activation that ensures that vinculin is activated only at sites of cell adhesion when two or more of its binding partners are brought into apposition.

We determined the structure of vinculin at 3.1 Å resolution in three different crystal forms grown from both high and moderate salt buffers. The tertiary and quaternary organization are identical in all cases. Vinculin comprises a ‘bundle of bundles’, with overall dimensions of 100 × 100 × 50 Å (Fig. 1 and Supplementary Fig. S1). The molecular building blocks are eight four-helix bundles, which are further organized into four tandem pairs, each of which resembles the structure of an amino-terminal fragment of α-catenin. We therefore call this class of domain a ‘vinculin/α-catenin repeat’ (VCR). Domains 1–3 comprise a VCR, in each of which a central helix runs the length (~100 Å) of the domain. In domains 1 and 2, the two four-helix bundles pack end to end. Domain 3 is more divergent, with the two bundles interacting along their sides and with a distinct break in helicity in the middle of the central helix. Domain 4 is a four-helix bundle that packs end to end with domain 5, which is the vinculin tail domain (Vt); together, these two domains topologically and structurally resemble a fourth VCR, but with a large insertion at the centre comprising the proline-rich region and the first helix of Vt. This repetitive tandem organization suggests that the molecule has arisen through gene duplication.

Domains 1, 2 and 3 together form a scaffold that buries a total of 5,000 Å² of surface area (Table 1) and forms the vinculin head (VH) observed in electron microscopy images⁶ (note that VH is distinct from the proteolytic fragment ‘Vh’, which includes D4 and part of the proline-rich region). D1 and D3 form a pair of pincers across which sits D2, locking them in position. The pincers hold Vt in the autoinhibited structure; the conformation of Vt is essentially identical to that seen in the crystal structure of the isolated tail⁷. The D1–D3 contact at the back of the pincers buries 1,560 Å² of surface area in a classic protein–protein contact with a hydrophobic core and polar or charged interactions at the periphery. The D2–D3 contact buries the most surface area (2,350 Å²), although it is more polar than the D1–D3 interface.

The N-terminal lobe of D2 (D2a) makes extensive contacts with D3, packing at right angles; the C-terminal lobe, D2b, is more or less

free of contacts with other domains (and is not present in *Drosophila* and *Caenorhabditis elegans* vinculins). The contacts between D4 and D3, although burying 1,050 Å² of surface, are largely polar and likely to be weak, suggesting that substantial interdomain movement is possible once Vt is freed from the head. Structural comparisons with the ‘M fragment’ of α-catenin⁸ further support this notion (see below). In electron microscopy images of the open conformation⁶, a ‘neck’ region was appended to the head with variable conformations, and we suggest that this neck is D4.

Vt makes two contacts with the vinculin head and one with the neck (Figs 1 and 2). As expected, D1 packs against Vt to form a major interface (2,290 Å² buried surface). In the crystal structure of a D1–Vt fragment⁹, the relative orientation of the two domains is essentially identical to that observed here, suggesting that the interface is rigid. The principal differences are found at the N terminus and the D1–D2 linker, where the linker adopts a non-native conformation in the D1–Vt fragment and three N-terminal residues take a different course. Two further intramolecular contacts contribute to the higher affinity of Vt for Vh than for D1. The first is a small end-to-end contact (~530 Å²) between the bottom of the Vt bundle and the top of D4; this contact includes main chain hydrogen bonding and a well-ordered salt bridge between Glu 775 and Arg 978. A second interface occurs on the opposite side to D1, where D3 is brought into close apposition with Vt; this interface buries 770 Å² and is largely polar.

Homology between the N-terminal VCR domains and the C-terminal tail domains of vinculin and α-catenin was previously established. Our structure further shows that the two domains of the M fragment of α-catenin are highly homologous to vinculin D3b and D4 (root-mean-square (r.m.s.) deviation of less than 1 Å; Fig. 3). However, the quaternary disposition of these two domains is highly variable in crystal structures of α-catenin^{8,10} and very different from that in full-length vinculin (Fig. 3b), and we propose that they provide a model of the open conformation of vinculin.

D4 precedes the α-catenin tail and, as in vinculin, it is separated by a flexible linker. This leaves unaccounted only a predicted four-helix bundle, which can be readily modelled by homology to D3a of vinculin. It is further possible that α-catenin shares a similar quaternary organization with vinculin. Thus, although crystal structures of α-catenin domain 1 lack the first helix¹¹, sequence alignments clearly indicate that it is present in the intact molecule (Supplementary Fig. S1). By assuming a vinculin-like head-to-tail

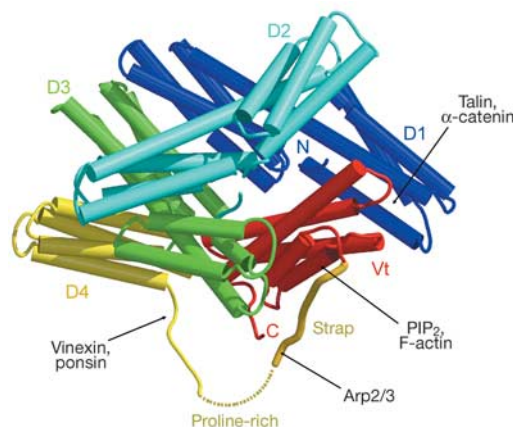


Figure 1 Structure of full-length vinculin in its autoinhibited state. Domains are shown in different colours: D1, residues 6–252; D2, 253–485; D3, 493–717; D4, 719–835; D5 (Vt), 896–1,066. A short N-terminal strand (‘N’, residues 1–5) precedes D1. The proline-rich region (838–878) is partly disordered and precedes a ‘strap’ (residues 878–890), that lies across the surface of Vt and the C terminus (‘C’). Binding sites for major ligands are indicated. PIP₂, PtdIns(4,5)P₂.

arrangement in α -catenin, we find that the D1–tail hydrophobic interface is highly conserved; and by assuming a vinculin-like D1–D3 interface, we also find a similar hydrophobic contact. Direct evidence for an autoinhibited conformation for α -catenin is currently limited, however, and dimerization also has a regulatory role¹⁰.

We previously showed that Vt undergoes a conformational change on binding phosphatidylinositol-4,5-bisphosphate (PtdIns(4,5)P₂), an acidic lipid that is upregulated at cell adhesion sites, and we identified two basic surfaces on Vt as plausible PtdIns(4,5)P₂-binding regions: a basic ‘collar’ surrounding the C-terminal arm, and a basic ‘ladder’ along an edge of helix 3 (ref. 7). Accordingly, point mutants in the collar (Lys911Ala and Lys924Ala) or basic ladder (Lys952Ala) reduce binding by 50%, whereas mutation of two basic residues in the neighbouring helix has no effect (Supplementary Fig. S2). Lys 911 and Lys 924 are next to His 906, a residue that is essential for PtdIns(4,5)P₂-induced conformational changes in Vt¹². In the context of the full-length molecule (Fig. 2), both the collar and the ladder point towards the outside of the molecule. The ladder is mostly exposed to solvent, although at its N-terminal end Lys 944 and Arg 945 make salt bridges to acidic residues on the head. The collar is partially obscured by a loop positioned N-terminal to the tail domain (the ‘strap’), which lies across the face of the H1–H2 hairpin. Thus, the quaternary organization of the closed conformation partly occludes the PtdIns(4,5)P₂-binding surface, as well as inhibiting PtdIns(4,5)P₂-dependent conformational changes in the tail.

At the bottom of the strap (which is the end of the proline-rich region) is a PPPP motif that is essential for binding to the Arp2/3 complex³. One of these prolines, Pro 878, packs against the base of Vt and the C-terminal arm, consistent with the low affinity of vinculin in its autoinhibited conformation for Arp2/3. We predict that on binding PtdIns(4,5)P₂ the strap is released from its location, which would rationalize the PtdIns(4,5)P₂ dependence of Arp2/3 binding³. At the beginning of the proline-rich region, an FPPPP motif is implicated in binding to VASP and vinexin^{13,14}. VASP binding is stimulated by PtdIns(4,5)P₂ (ref. 15), suggesting that in the isolated molecule this region packs against the molecular surface. In our structure, however, this region is involved in a crystal contact and it is unclear what its conformation would be in the isolated molecule. The rest of the proline-rich region is poorly

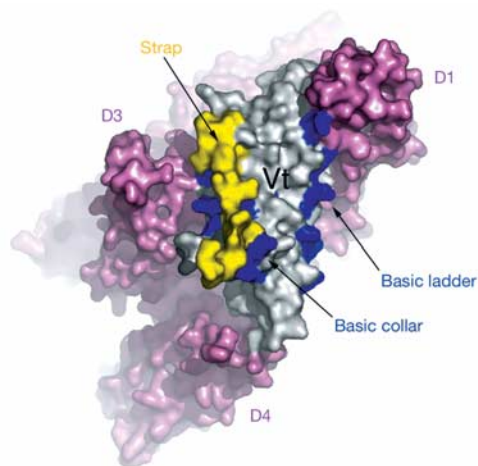


Figure 2 Accessible surface of Vt in the context of the full-length molecule. Vt (grey) is surrounded on three sides by D1, D3 and D4. Basic residues comprising two regions (the basic ladder and basic collar) on Vt implicated in PtdIns(4,5)P₂ binding are shown in blue. The strap (yellow) partly occludes the basic collar, whereas the basic ladder is mostly accessible, except at the top where it contacts D1. Behind the strap lies His 906, a residue implicated in triggering PtdIns(4,5)P₂-dependent conformational changes in Vt.

ordered or variable in conformation and does not pack against the molecular surface, consistent with the sensitivity of this region to V8 protease cleavage¹⁶.

Talin rod does not bind to full-length vinculin in solution (Fig. 4b). The binding of talin domains to vinculin D1 is strong but characterized by slow kinetics, suggestive of a conformational change on binding¹⁷. To test this hypothesis, we expressed a fragment of vinculin corresponding to domains D1–D3 (VH, residues 1–718). In the absence of a binding partner D1 is loosely folded, as judged by its low melting temperature and enthalpy of unfolding (Fig. 4c). On binding Vt, the VH structure becomes highly ordered (Fig. 4d). Binding of the talin rod (relative molecular mass 220,000; *M_r* 220K) to VH results in the appearance of a similar strong new peak (Fig. 4e), consistent with tight binding to the D1 domain (dissociation constant, *K_d* ≈ 40 nM)¹⁷.

As the binding site for α -catenin on VH had not previously been mapped, we expressed a domain-sized fragment (equivalent to D3 of vinculin) of α -catenin containing the crucial region for vinculin binding¹⁸: this fragment binds to vinculin D1 in a very similar fashion to the binding of talin rod (Fig. 4f). Using isothermal titration calorimetry (ITC), we observed tight binding (*K_d* = 82 ± 19 nM) that was endothermic (ΔH = +8.2 ± 1.2 kcal mol^{−1}) and entropy driven (+60 ± 5 cal K^{−1} mol^{−1}), consistent with a major conformational change on binding (Supplementary Table S1).

To explore this further, we designed a mutant that would stabilize the conformation of D1 observed in our vinculin crystal structure. The mutant Ala50Ile fills a cavity in the hydrophobic core between helices H1 and H2 (Fig. 4a). Melting profiles show that the Ala50Ile mutation in the context of VH does indeed stabilize the D1 fold to a significant extent (Fig. 4c); moreover, Vt binds more strongly to the mutant than to the wild-type VH (Fig. 4d). However, binding of the mutant VH to α -catenin is reduced by a factor of about 10 (*K_d* = 710 ± 40 nM) and talin binding is also greatly reduced (Fig. 4e, f). Thus, binding of both talin and α -catenin to VH

Table 1 Buried interfacial area between domains (Å²)

	D2	D3	D4	D5 (Vt)
D1	1,240	1,560	0	2,290
D2		2,350	0	30
D3			1,050	770
D4				540

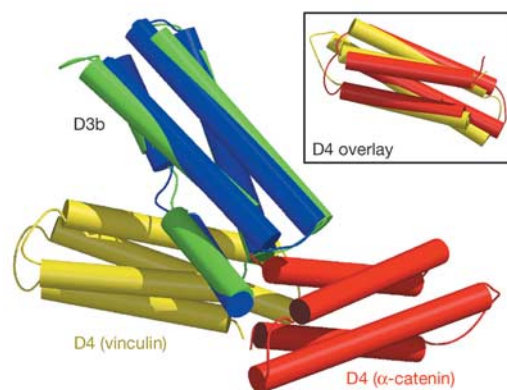


Figure 3 Comparison of vinculin domains 3b and 4 with the ‘M fragment’ of α -catenin. Overlay is on the D3b domains (r.m.s. deviation 0.79 Å for 110 C α atoms), which shows a roughly 180° rotation of the D4 domains. The view and colouring for vinculin are the same as in Fig. 1. Inset shows an independent overlay of the D4 domains (r.m.s. deviation 0.86 Å for 115 C α atoms).

requires conformational changes in the D1 bundle that are inhibited by Vt binding. A crystal structure of D1 in complex with a short talin-derived peptide supports this hypothesis⁹.

Head-to-tail binding has been estimated with proteolytic and recombinant fragments ($K_d \approx 20$ –50 nM), but this value presumably underestimates the intramolecular affinity when the head and tail are covalently linked. A peptide derived from the talin rod (residues 1,944–1,969) activates full-length vinculin when present in a 500–1,000-fold molar excess by binding to vinculin D1 (ref. 17). We measured the binding of this peptide to VH by ITC ($K_d = 400 \pm 150$ nM). Assuming that binding to full-length vinculin is negligible and that the system has achieved equilibrium, this gives an estimate for the intramolecular VH–Vt K_d of less than 1 nM. A similar result is obtained for a phage-display peptide that binds to the head with an affinity of 100 nM and activates vinculin when present in a 300-fold excess, implying an intramolecular K_d of about 0.3 nM (refs 19, 20). These values imply an affinity that is roughly 100 times higher than that for the isolated VH and Vt domains, explaining why talin rod ($K_d \approx 30$ nM for binding to the isolated head) and α -catenin ($K_d = 80$ nM) do not bind significantly to full-length vinculin when present at an equimolar concentration.

Our data show that the binding sites on the head, tail and proline-rich regions are structurally distinct, but conformationally and thus thermodynamically linked. Therefore, activation of vinculin could be achieved by a combinatorial input of ligands to these distinct regions in which the activating potentials (that is, the difference in energy between binding to the closed and open states) are essentially additive. Once these potentials exceed the intramolecular binding energy, the molecule should open and binding should ensue,

provided that a kinetic pathway to equilibrium is available. Estimates of phospholipid (10% PtdIns(4,5) P_2 in phosphatidylcholine vesicles) binding to Vt are in the micromolar range, whereas binding to the full-length molecule is negligible at physiological salt²¹. This binding energy is insufficient to activate vinculin by itself, but in combination with PtdIns(4,5) P_2 micelles (in a 100-fold molar excess) talin forms a ternary complex with vinculin and PtdIns(4,5) P_2 (ref. 22). Because the PtdIns(4,5) P_2 -binding site is only partly occluded in the full-length molecule, we can propose a kinetic pathway to activation in which PtdIns(4,5) P_2 binding and the ensuing conformational changes transiently release the tail from the head, allowing talin or α -catenin to bind to the head and ligands such as VASP and vinexin to bind the proline-rich region. The full repertoire of input signals remains to be determined but may also include the binding of tensin to the neck (D4; D. Lin, personal communication).

Similar considerations apply to the binding of F-actin, where the binding affinity between F-actin and Vt in physiological salt has been estimated at $K_d \approx 1 \mu\text{M}$, whereas binding to the full-length protein is very weak. In the structure of Vt in complex with F-actin²³, the actin-binding surface comprises two regions of Vt. One of the binding surfaces is exposed in the full-length molecule and is available for binding. The second site is only partly exposed and is blocked by the vinculin head. The principal steric clash occurs with the end of helix 1 in D1, which has been implicated in structural rearrangements on binding talin⁹. This arrangement provides a kinetic pathway for F-actin to bind to vinculin and to transiently loosen intramolecular interactions, leading to activation if a head-binding ligand is available. Because F-actin binding and PtdIns(4,5) P_2 binding are exclusive²⁰, these ligands provide two alternative pathways to activation.

The organization of tandem domains into an autoinhibited conformation that provides a combinatorial output to numerous input signals is an emerging theme among signal transduction proteins^{24,25}. Vinculin extends this theme to proteins that lack a catalytic domain, that seem to have arisen from gene duplication of a single protein interaction domain, and that respond primarily to the spatial colocalization of their binding partners rather than to posttranslational modification. □

Methods

Protein expression and purification

Recombinant His-tagged chicken vinculin was cloned into a pET15b vector, expressed in *Escherichia coli* BL21(DE3) cells, and purified with a HiTrap affinity column (Amersham Pharmacia Biotech). After dialysis against 20 mM sodium phosphate (pH 7.5), 150 mM NaCl and 0.1 mM EDTA, the protein was further purified by ion exchange chromatography (DE52) and concentrated in an ultrafiltration cell (Amicon) to 6 mg ml⁻¹. Wild-type and Ala501le mutant vinculin D1–D3 (residues 1–718) and α -catenin D3 (residues 273–510) were cloned, expressed and purified by a similar protocol. We expressed His-tagged talin rod (residues 397–2,541) in a pET30a vector and purified it as described²⁶. Vt was expressed and purified as described⁷. Wild-type vinculin was purified from chicken gizzard smooth muscle as described²⁷. Recombinant selenomethionine (SeMet) vinculin was expressed in minimal media supplemented with SeMet as described for Vt⁷.

Structure determination

Crystals of recombinant vinculin were grown at 30 °C by a hanging-drop vapour diffusion method, with protein at 4–6 mg ml⁻¹ in 20 mM NaCl, 20 mM Tris pH 8.0 and 5 mM dithiothreitol, against a reservoir of 100 mM cacodylate (pH 6.5) and 1–1.2 M (NH₄)₂SO₄. Three crystal forms were obtained (C2: $a = 57 \text{ \AA}$, $b = 126 \text{ \AA}$, $c = 170 \text{ \AA}$, $\beta = 94.9^\circ$; P2₁: $a = 57 \text{ \AA}$, $b = 353 \text{ \AA}$, $c = 69 \text{ \AA}$, $\beta = 114.1^\circ$; and C222₁: $a = 56 \text{ \AA}$, $b = 127 \text{ \AA}$, $c = 352 \text{ \AA}$). Crystals of chicken gizzard vinculin grew in 30% PEG 5000 MME, 0.2 M (NH₄)₂SO₄, 100 mM MES pH 6.5 and 10 mM MgSO₄, and were very similar to the orthorhombic recombinant crystals, although they diffracted less well ($d_{\text{min}} 3.9 \text{ \AA}$). Crystals were transferred directly to a cryobuffer consisting of 1.5–1.8 M (NH₄)₂SO₄, 100 mM cacodylic acid (pH 6.5) and 20% glycerol, flash-frozen in an N₂ stream, and then annealed by re-equilibration in cryobuffer for a few minutes before being returned to the N₂ stream.

Multiwavelength anomalous diffraction data were collected from orthorhombic SeMet crystals at the Stanford Synchrotron Radiation Laboratory (SSRL) beamline 9.1 (Supplementary Table S2). Crystal lifetime limited data collection to two wavelengths: the high-energy remote and inflection point (0.979 and 0.94 Å, respectively). Because an increase in mosaic spread and cell volume was observed, the data were reprocessed using

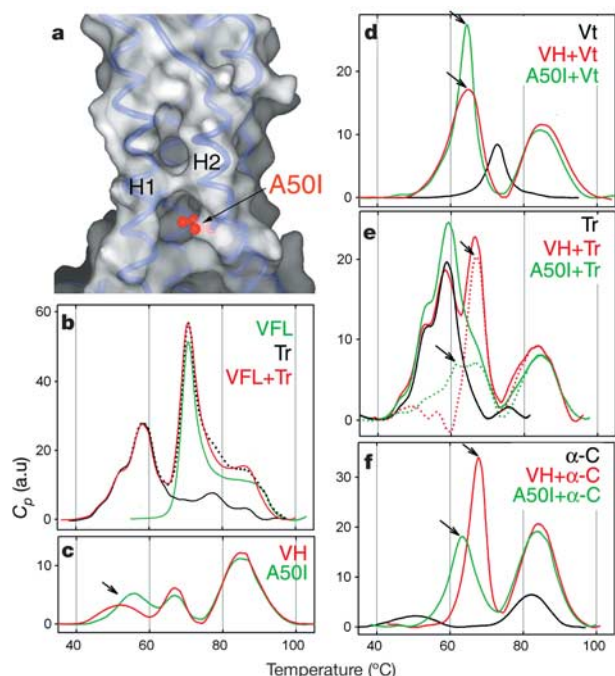


Figure 4 Calorimetric analysis of vinculin and ligand complexes. **a**, Surface representation of D1 showing the cavity that is filled by the Ala501le (A501I) mutation. Helices H1 and H2 are indicated and the mutation is shown in ball-and-stick notation. **b**, Melting profiles (in arbitrary units) of full-length vinculin (VFL) and talin rod (Tr) alone (13.0 μM) and after 1 h of incubation. Broken line shows the sum of component curves. **c**, Melting profile of VH and the Ala501le mutant. Arrow marks the peak ascribed to D1. **d–f**, Melting profiles of wild-type and mutant VH complexes with Vt (**d**), talin rod (**e**) and α -catenin D3 (α -C; **f**). Arrows mark the new peaks associated with complex formation. In **e**, broken lines indicate the difference profiles (complex minus talin rod alone).

local scaling. The anomalous component of the structure factor was calculated by XPREP, and Patterson correlation analysis was done by the program SHELXD²⁸ at a resolution of 4.5 Å. Because there was no clear occupancy drop for the sites, phase refinement was carried out with SHARP²⁹, which refined coordinates and occupancy but not temperature factors. By this approach, 25–27 selenium sites were confirmed out of an expected total of 39, leading to a map with visible helical density.

Phased molecular replacement using BRUTEPTF (<http://russell.bioc.aecom.yu.edu/server/NYSGR.html>) located the intact Vt (Protein Data Bank (PDB) accession code 1QKR) and α -catenin dimerization (PDB 1DOV) domains, as well as one half of the α -catenin M-fragment (PDB 1H6G). Chain tracing for the remainder of the molecule was guided by the selenium positions. Refinement was done with the CNS package³⁰ against the maximum-likelihood target using amplitudes and phase probability distribution to 3.1 Å resolution. No σ cut-off was used in refinement. Convergence was reached at an R_{work} of 31.6% and an R_{free} of 35.7%. The current model includes residues 1–855 and 875–1,065. Of these, 87.4%, 10.9% and 1.5% are in the most favoured, additionally and generously allowed regions, respectively, of the Ramachandran plot. The structure was used as a molecular replacement search model for the other crystal forms. Refinement proceeded for each to R_{free} values of 0.35–0.40. Minor differences were observed in interhelical loops and the proline-rich region at sites of crystal contacts. Coordinates for the model of α -catenin are available from the authors on request.

Calorimetry

Differential scanning calorimetry (DSC) experiments were done at a scanning rate of 1 K min^{−1} under 3.0 atm of pressure on an N-DSC II differential scanning calorimeter (Calorimetry Sciences). Before measurements, all protein samples were dialysed against PBS buffer, which was used as the reference solution. All thermal transitions were irreversible under the conditions used. See Supplementary Methods for a further description of the DSC method. ITC was carried out on a VP-ITC calorimeter (Microcal). We injected 8- μ l aliquots of solution containing either 940 μ M α -catenin D3 or 1.0 mM talin peptide into a cell containing 70–100 μ M vinculin D1–D3 construct (wild type or the Ala50Ile mutant). In each experiment 35 injections were made. The experiments were done at 23°C. Before ITC titrations, all protein samples were dialysed against PBS buffer. Experimental data were analysed with Microcal Origin software provided by the ITC manufacturer (Microcal).

Received 1 April; accepted 29 April 2004; doi:10.1038/nature02610.
Published online 13 June 2004.

- Volberg, T. *et al.* Focal adhesion formation by F9 embryonal carcinoma cells after vinculin gene disruption. *J. Cell Sci.* **108**, 2253–2260 (1995).
- Xu, W. M., Baribault, H. & Adamson, E. D. Vinculin knockout results in heart and brain defects during embryonic development. *Development* **125**, 327–337 (1998).
- DeMali, K. A., Barlow, C. A. & Burridge, K. Recruitment of the Arp2/3 complex to vinculin: coupling membrane protrusion to matrix adhesion. *J. Cell Biol.* **159**, 881–891 (2002).
- Johnson, R. P. & Craig, S. W. An intramolecular association between the head and tail domains of vinculin modulates talin binding. *J. Biol. Chem.* **269**, 12611–12619 (1994).
- Johnson, R. P. & Craig, S. W. F-actin binding site masked by the intramolecular association of vinculin head and tail domains. *Nature* **373**, 261–264 (1995).
- Winkler, J., Lunsdorf, H. & Jockusch, B. M. The ultrastructure of chicken gizzard vinculin as visualized by high-resolution electron microscopy. *J. Struct. Biol.* **116**, 270–277 (1996).
- Bakolitsa, C., de Pereda, J. M., Bagshaw, C. R., Critchley, D. R. & Liddington, R. C. Crystal structure of the vinculin tail suggests a pathway for activation. *Cell* **99**, 603–613 (1999).
- Yang, J., Dokurno, P., Tonks, N. K. & Barford, D. Crystal structure of the M-fragment of α -catenin: implications for modulation of cell adhesion. *EMBO J.* **20**, 3645–3656 (2001).
- Izard, T. *et al.* Vinculin activation by talin through helical bundle conversion. *Nature* **427**, 171–175 (2004).
- Pokutta, S., Drees, F., Takai, Y., Nelson, W. J. & Weis, W. I. Biochemical and structural definition of the I-afadin- and actin-binding sites of α -catenin. *J. Biol. Chem.* **277**, 18868–18874 (2002).
- Pokutta, S. & Weis, W. I. Structure of the dimerization and β -catenin-binding region of α -catenin. *Mol. Cell* **5**, 533–543 (2000).
- Miller, G. J. & Ball, E. H. Conformational change in the vinculin C-terminal depends on a critical histidine residue (His-906). *J. Biol. Chem.* **276**, 28829–28834 (2001).
- Brindle, N. P. J., Holt, M. R., Davies, J. E., Price, C. J. & Critchley, D. R. The focal adhesion vasodilator-stimulated phosphoprotein (VASP) binds to the proline-rich domain in vinculin. *Biochem. J.* **318**, 753–757 (1996).
- Kioka, N. *et al.* Vinexin: a novel vinculin-binding protein with multiple SH3 domains enhances actin cytoskeletal organisation. *J. Cell Biol.* **144**, 59–69 (1999).
- Huttelmaier, S. *et al.* The interaction of the cell-contact proteins VASP and vinculin is regulated by phosphatidylinositol-4,5-bisphosphate. *Curr. Biol.* **8**, 479–488 (1998).
- Price, G. J. *et al.* Primary sequence and domain-structure of chicken vinculin. *Biochem. J.* **259**, 453–461 (1989).
- Bass, M. D. *et al.* Further characterization of the interaction between the cytoskeletal proteins talin and vinculin. *Biochem. J.* **362**, 761–768 (2002).
- Imamura, Y., Itoh, M., Maeno, Y., Tsukita, S. & Nagafuchi, A. Functional domains of α -catenin required for the strong state of cadherin-based cell adhesion. *J. Cell Biol.* **144**, 1311–1322 (1999).
- Adey, N. B. & Kay, B. K. Isolation of peptides from phage-displayed random peptide libraries that interact with the talin-binding domain of vinculin. *Biochem. J.* **324**, 523–528 (1997).
- Steinle, P. A., Hoffert, J. D., Adey, N. B. & Craig, S. W. Polyphosphoinositides inhibit the interaction of vinculin with actin filaments. *J. Biol. Chem.* **274**, 18414–18420 (1999).
- Johnson, R. P., Niggli, V., Durrer, P. & Craig, S. W. A conserved motif in the tail domain of vinculin mediates association with and insertion into acidic phospholipid bilayers. *Biochemistry* **37**, 10211–10222 (1998).
- Gilmore, A. P. & Burridge, K. Regulation of vinculin binding to talin and actin by phosphatidylinositol-4-5-bisphosphate. *Nature* **381**, 531–535 (1996).

- Janssen, M. E. W. *et al.* Three-dimensional structure of vinculin bound to actin filaments. *J. Cell Biol.* (submitted).
- Liddington, R. C. & Bankston, L. A. The structural basis of dynamic cell adhesion: heads, tails and allostery. *Exp. Cell Res.* **261**, 37–43 (2000).
- Lim, W. A. The modular logic of signaling proteins: building allosteric switches from simple binding domains. *Curr. Opin. Struct. Biol.* **12**, 61–68 (2002).
- Xing, B., Jedsadayamata, A. & Lam, S. C. Localization of an integrin binding site to the C terminus of talin. *J. Biol. Chem.* **276**, 44373–44378 (2001).
- O'Halloran, T., Molony, L. & Burridge, K. Purification and assay of vinculin, metavinculin, and talin. *Methods Enzymol.* **134**, 69–77 (1986).
- Uson, I. & Sheldrick, G. M. Advances in direct methods for protein crystallography. *Curr. Opin. Struct. Biol.* **9**, 642–648 (1999).
- de la Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276**, 472–494 (1997).
- Brunker, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the Cell Migration Consortium, the NIH, the BBSRC and the Wellcome Trust for financial support; the DOE and staff at SSRL for synchrotron access and support; S. Lam for the talin rod expression vector; M. Zhang for talin expression; and B. Patel for vinculin purification. D.M.C. is supported by a Howard Hughes predoctoral fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.C.L. (rlidding@burnham.org). The atomic coordinates have been deposited in the PDB under accession codes 1ST6 (coordinates) and r1ST6sf (structure factors).

Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR

Dmitry M. Korzhnev^{1,2*}, Xavier Salvatella^{3*}, Michele Vendruscolo³, Ariel A. Di Nardo¹, Alan R. Davidson¹, Christopher M. Dobson³ & Lewis E. Kay^{1,2}

¹Departments of Molecular and Medical Genetics and Biochemistry, University of Toronto, Toronto, Ontario M5S 1A8, Canada

²Department of Chemistry and Protein Engineering Network Center of Excellence, University of Toronto, Toronto, Ontario M5S 1A8, Canada

³Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

* These authors contributed equally to this work

Many biochemical processes proceed through the formation of functionally significant intermediates^{1,2}. Although the identification and characterization of such species can provide vital clues about the mechanisms of the reactions involved, it is challenging to obtain information of this type in cases where the intermediates are transient or present only at low population^{1–4}. One important example of such a situation involves the folding behaviour of small proteins that represents a model for the acquisition of functional structure in biology¹. Here we use relaxation dispersion nuclear magnetic resonance (NMR) spectroscopy to identify, for two mutational variants of one such protein, the SH3 domain from Fyn tyrosine kinase⁵, a low-population folding intermediate in equilibrium with its unfolded and fully folded states. By performing the NMR experiments at different temperatures, this approach has enabled characterization of the kinetics and energetics of the folding process as well as providing structures of the intermediates. A general strategy emerges for an experimental determination of the energy landscape of a protein by applying this methodology to a series of

mutants whose intermediates have differing degrees of native-like structure.

Investigating the folding of small single-domain proteins has been crucial in understanding how biological molecules have evolved to self-assemble into complex functional structures^{6–8}. SH3 domains are protein modules consisting of about 60 residues that fold into five-stranded β -sandwiches composed of two orthogonal β -sheets⁹ (see Fig. 1a). The folding behaviour of several of these domains has been characterized in detail and found to be consistent with a two-state process; that is, there is no measurable population of partly structured intermediates^{9,10}. Indeed, studies of this behaviour have provided important insights into the detailed mechanism of protein folding^{11,12}.

We have recently shown that mutations of the highly conserved glycine residue at position 48 of the Fyn SH3 domain both destabilize the protein and markedly increase the folding rate¹³, k_f . We now use ¹⁵N relaxation dispersion NMR spectroscopy¹⁴ to probe the folding/unfolding reaction of two Fyn SH3 mutants, G48M and G48V, under native conditions. In these experiments, contributions to line widths of cross peaks in ¹H–¹⁵N correlation spectra are interpreted in terms of the rates of interconversion between states (kinetics), their chemical-shift differences (structure) and their populations (thermodynamics)^{14,15}. For exchange rates varying from a few hundred to a few thousand per second and, for populations of the higher-energy (unfolded or intermediate) state greater than about 0.5%, this approach provides a sensitive measure of the exchange dynamics. Distinct advantages of the NMR experiments over global measurements such as those provided by circular dichroism or fluorescence are that site-specific folding information is available from each probe (in this case individual backbone amides) and that the experiments can be performed in the absence of denaturant.

Figure 1b, c (insets) illustrates typical ¹⁵N Carr–Purcell–Meiboom–Gill (CPMG) dispersion profiles for Ala 12, a representative residue of Fyn SH3, analysed under the assumption of a two-state exchange process (solid lines). The kinetic parameters obtained are very similar to those extrapolated from unfolding/folding experiments in urea by stopped-flow fluorescence¹³ (see Supplementary Table 1S), strongly supporting our interpretation that the increased line widths in NMR spectra arise from the folding reaction.

The high quality of the fits and the associated statistical analysis show that more complex models of exchange are not required to explain the experimental data when dispersion profiles are considered independently for the different residues. However, when k_f and unfolding rate (k_u) values measured for different probes (namely, different amides) are compared (Fig. 1b, c), it becomes clear that the exchange process is more complex for both mutants, particularly at temperatures below 20 °C. In the case of two-state folding, all of the folding/unfolding rates obtained should be the same within experimental error, because all probes are reporting on the same physical process. By contrast, if the folding mechanism is more complex, fits to a two-state exchange model are expected to produce residue-specific variations in the parameters describing the exchange process. The latter is true for many residues of the SH3 mutants at lower temperatures (Fig. 1b, c) with the differences in observed rates significantly larger than the error bounds. However, it is interesting to note that k_f (k_u) values for different residues converge with increasing temperature, so that the unfolding or folding process approximates to a two-state model at temperatures above about 25 °C for both G48M and G48V.

Recognizing that the exchange kinetics are more complex than two-state under at least some conditions, the dispersion data at all temperatures and for a significant set of residues with large dispersions were fitted simultaneously to a three-state exchange model that includes U, I and F states (see Methods). In this approach we assume that the chemical-shift differences between states are invar-

iant with temperature, an assumption supported by previous work of Palmer *et al.*¹⁴, that the temperature dependence of the rates obeys standard transition-state theory¹ and that over the range of temperatures examined ΔH and ΔS are essentially constant (see Methods). Figure 2a, b shows examples of fits of the dispersions, here for Ser 41 and Thr 44 of the G48M mutant, based on the global analysis.

The temperature dependence of the populations of U, I and F, along with their rates of interconversion, are shown in Fig. 2c, d. The population of the I state is relatively independent of temperature, whereas that of the U state increases from 3% to 5% and from 1% to

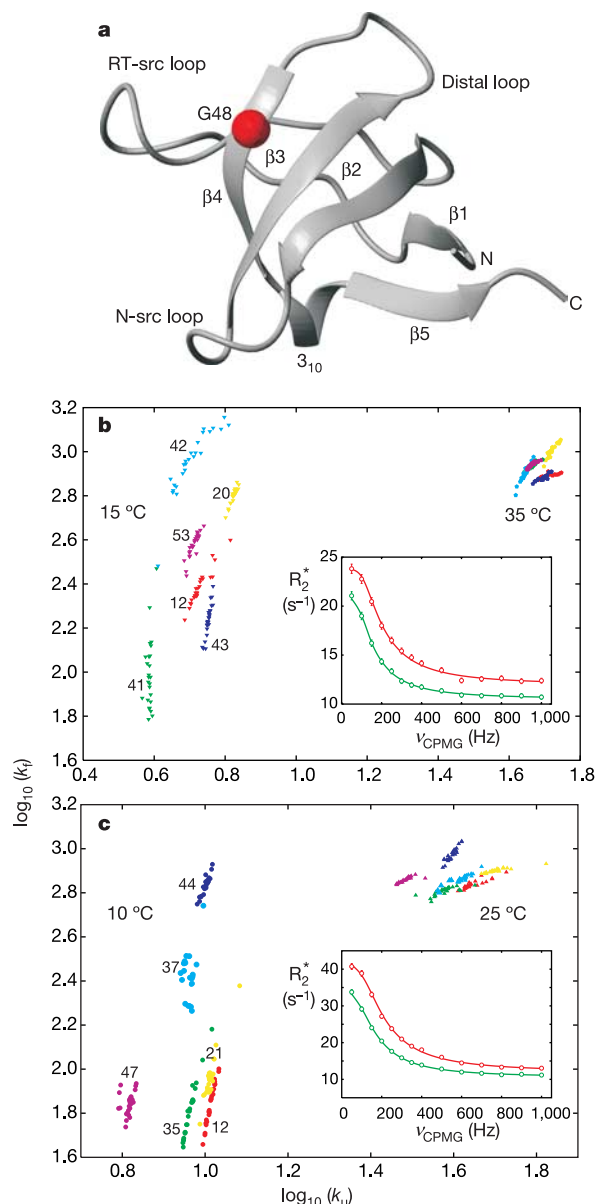


Figure 1 Measurement of k_f and k_u in the G48M and G48V Fyn SH3 domains. **a**, Ribbon diagram of the wild-type Fyn SH3 domain (pdb accession code 1SHF; ref. 27), highlighting the site of mutation, Gly 48. **b**, **c**, Values of k_f and k_u for selected residues (labelled with residue number) from fits of dispersions assuming a two-state model at 15 and 35 °C for G48M (**b**) and 10 and 25 °C (G48V) (**c**). Errors in rates were measured with a jackknife procedure by which 25% of the points in each dispersion were randomly removed; subsequent fits of the data produce the ranges of rates indicated. Insets show representative dispersion curves (for Ala 12) recorded at 800 MHz (red) and 600 MHz (green) at 25 °C, with R_2^* the measured relaxation rate and ν_{CPMG} the field strength employed.

5% for G48V and G48M, respectively, over the temperature range examined. In Fig. 2d we have plotted $k_{UI} + k_{IU}$ (labelled UI, where k_{ij} is the rate from state i to j) and $k_{IF} + k_{FI}$ (labelled IF) for both G48V and G48M. In particular, $k_{UI} + k_{IU}$ increases rapidly with temperature so that at higher temperatures the exchange between states U and I approaches the fast regime on the NMR chemical-shift timescale for both mutants, leading to a rapid equilibration between U and I.

This analysis allows relative values to be extracted for the thermodynamic parameters of each state for each mutant, yielding the results shown in Fig. 2c, where the U state is used as a reference (see Supplementary Table 3S). The G48M mutant is more stable than G48V, ΔG_{U-F} ($G_U - G_F$) = 2.3 kcal mol⁻¹ (1 kcal ≈ 4.18 kJ) versus 1.8 kcal mol⁻¹, with the I state similar in free energy to the U state (within 0.5 kcal mol⁻¹ in both cases). Notably, very similar ΔG_{U-F} values are obtained by fitting the data to a two-state model (within 0.2 kcal mol⁻¹), whereas the ΔH_{U-F} values extracted from the two models (two-state versus three-state) differ by less than about 25%. The potential interplay between H and S values¹⁶, and the difficulties involved in deconvolving the contributions of the

thermodynamic properties of the polypeptide chain from those of the solvent, means that their profiles must be interpreted more conservatively than those for G. Nevertheless, for both mutants $H_I \ll H_U$, an observation that is compatible with the formation of considerable structure in the I state. However, the larger enthalpy difference between states I and F for G48M ($\Delta H_{F-I} = -3.4$ kcal mol⁻¹) than for G48V ($\Delta H_{F-I} = 0.7$ kcal mol⁻¹) is particularly intriguing and indicates that the I state of G48V might be structurally more native-like than that of G48M (see below). In addition, this analysis indicates that the relative contribution of the entropic term to the free energy barrier for folding of Fyn SH3 is rather small, indicating that the expected decrease in chain entropy could be offset by the increased disorder of the solvent, particularly as the folding rates were measured at relatively low temperatures^{1,17,18}.

In addition to a kinetic and thermodynamic characterization of the three-site exchange process, structural information is also available from the dispersion experiments in the form of backbone ¹⁵N chemical-shift differences between the various states. Figure 3a shows these differences, displayed in the form of a ratio,

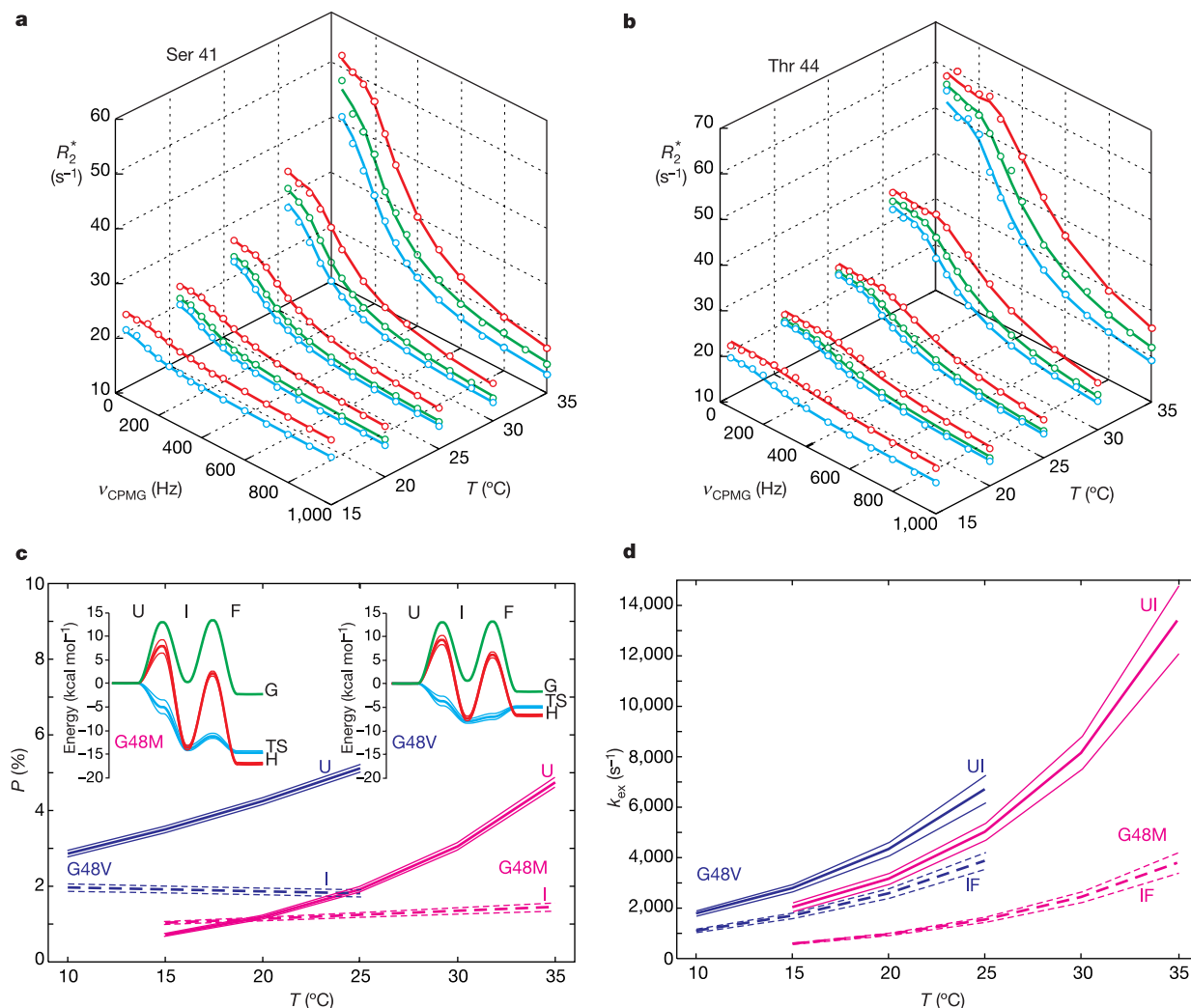


Figure 2 Three-site folding model for G48M and G48V Fyn SH3. **a, b**, Typical fits of dispersion data with a global three-state folding model for Ser 41 (**a**) and Thr 44 (**b**) of the G48M Fyn SH3 domain. Data from all residues, all temperatures and all spectrometer fields (blue, 500 MHz; green, 600 MHz; red, 800 MHz) were included in the analysis. **c, d**, Temperature dependence of populations (**c**) and rates of interconversion between states (**d**), along with energy diagrams for G48M (25 °C) and G48V (17.5 °C) are shown in

the middle curve of each three, with the outer curves corresponding to errors of 1 standard deviation, estimated by the covariance matrix method²⁸. The k_{ex} rates labelled UI and IF correspond to $k_{UI} + k_{IU}$ and $k_{IF} + k_{FI}$, respectively. Note that placement of the reference state (U) of each mutant in the energy-level diagrams is arbitrary; comparison of absolute parameters between mutants must therefore be made with caution.

$|\Delta^{\text{exp}}(k)| = |\delta_{\text{FI}}(k)/\delta_{\text{FU}}(k)|$, colour coded on the structure of the wild-type protein, where $\delta_{ij}(k)$ is the difference in ^{15}N shifts between states i and j for residue k . For a given residue, a ratio smaller than 0.5 indicates that the I state shift is closer to that of the folded state than to that of the unfolded state. An important validation of the analysis is the close similarity of the chemical-shift data for the U states of the two mutants (all chemical-shift data are provided in Supplementary Information). However, the chemical shifts of the I state in G48V seem on average to be significantly more native-like than the corresponding shifts in the G48M mutant

($\langle\Delta^{\text{exp}}\rangle_{\text{G48M}} = 0.61$, whereas $\langle\Delta^{\text{exp}}\rangle_{\text{G48V}} = 0.38$).

The measured chemical-shift differences were then used to calculate structural ensembles for the I state by using an approach similar to that originally developed to determine transition-state structures from protein engineering ϕ values¹⁹. Because most of the conformational factors that are known to influence ^{15}N shifts in native proteins can be related to the formation of inter-atomic interactions²⁰, in this approach we assume that the degree of unfolding of a given residue is related to Δ^{exp} , and computationally to the fraction of native contacts lost by the amide nitrogen of that residue in the I state, Δ^{calc} . A simulated annealing protocol developed to sample protein configurations for which $\Delta^{\text{exp}}(k) = \Delta^{\text{calc}}(k)$ (Fig. 3b) was used to produce an ensemble of 25 structures for the I state of each Fyn SH3 mutant.

The ensembles of conformations (Fig. 3c) and the ensemble-averaged energy maps (Fig. 3d) illustrate how in both intermediates the topology of the native state is already defined. Analysis of the calculated structures indicates that the most native-like region of both intermediates is the β -sheet formed by strands 2, 3 and 4 (the average root-mean-square deviation (r.m.s.d.) of residues in strand 4 between the intermediates and the native state is 3.7 and 1.9 Å for mutants G48M and G48V, respectively) and that their only significant structural differences with the native state are found in the connecting regions such as the RT loop and at the termini of the protein (details given in Supplementary Fig. 1S). Interestingly, several studies on the folding of Fyn SH3 domains have clearly established that the region of the native protein involving residues 45–50, which is part of strand 4, is formed almost completely in the transition state^{9,11,12,21}. Given the fundamental differences between the two techniques used to probe the structure present in these states (ϕ values for transition states and ^{15}N chemical shifts for intermediates) it is remarkable that similar structural regions are identified in each case. However, the I states of the G48M and G48V mutants differ significantly in their structural heterogeneity (the average r.m.s.d. between two members of the ensemble is 5.6 Å in G48M and 3.3 Å in G48V) and in their degree of compaction. A recent study has shown a very strong correlation of folding kinetics with both β -sheet propensity and residue volume in a series of mutations at residue 48 (ref. 13); taken together, these results show that, although the acquisition of the native topology is the dominant structural event in the folding of Fyn SH3, the interactions established by the side chain at position 48 are very important in defining both the energetics and the degree of structure present at different stages along the folding reaction.

Thus, we have introduced here a novel methodology for the study of the mechanisms of protein folding and have applied it to characterize the behaviour of two destabilized, but fast-folding, mutants of the Fyn SH3 domain. Despite the apparent two-state nature of the folding and unfolding of this protein, we have shown that an intermediate can be characterized even though it is present at a level of only about 1% of the total population of molecules; because the folding of Fyn SH3 seems to be typical of small proteins, we believe that this approach will be widely applicable. Our results indicate that similar regions of the overall structure are present in the G48M and G48V intermediates, but also that these states differ in their degree of structural homogeneity and compaction owing to the crucial role of position 48 in the folding of Fyn SH3 (ref. 13). The two mutants that we have studied have allowed different stages of the folding landscape to be explored, indicating that a more complete picture could emerge from the investigation of a wider range of mutants whose intermediates are likely to span a greater portion of conformational space. This study is therefore an important step forwards in characterizing the detailed energy landscape of a folding reaction through the generation of extensive kinetic, thermodynamic and structural information. □

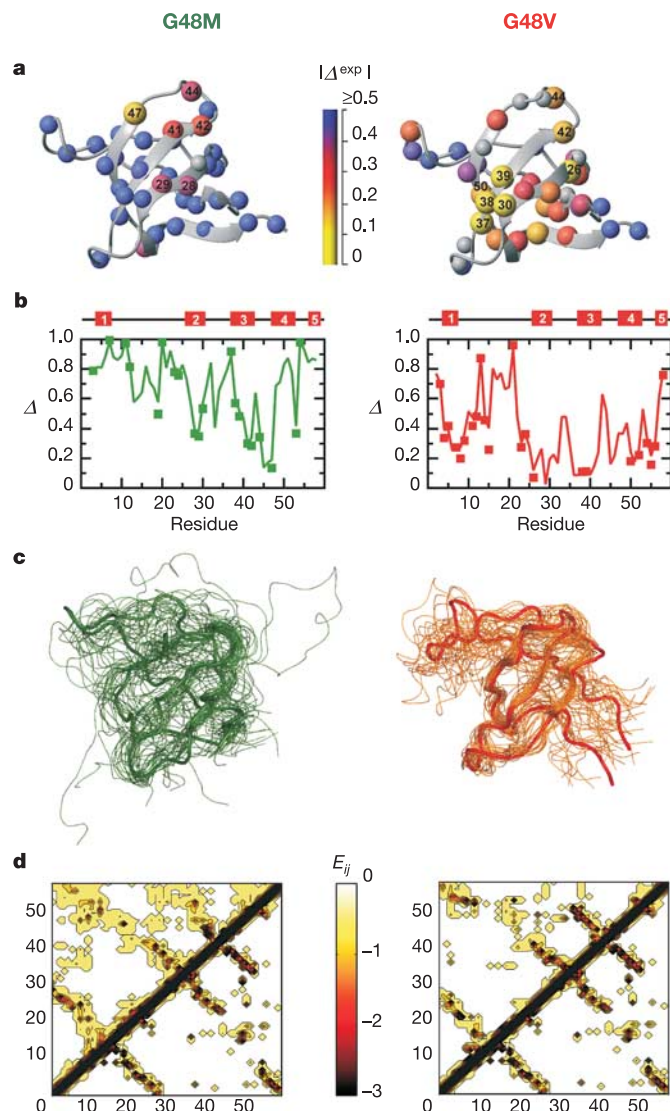


Figure 3 Structural analysis of the I state of mutants G48M and G48V of Fyn SH3 using chemical shifts. **a**, Ratio of ^{15}N chemical-shift differences, $|\Delta^{\text{exp}}(k)| = |\delta_{\text{FI}}(k)/\delta_{\text{FU}}(k)|$, plotted on the ribbon trace of the structure of the wild-type Fyn SH3 domain²⁷, with the colour code indicated. Residues with values of $\Delta^{\text{exp}} < -0.5$ or $\Delta^{\text{exp}} > 2.5$ are in grey. **b**, Comparison of Δ^{exp} (filled squares) with the Δ^{calc} values obtained in the structure determinations (lines); only residues with more than four native contacts have been represented. **c**, Ribbon representation of the ensemble of 25 structures determined for each mutant; the structure with the thicker ribbon is the more representative (lowest average pairwise r.m.s.d.) of the ensemble. **d**, Average effective energy contact map of the native state (bottom) and the ensembles obtained for the I state of each mutant (top); E_{ij} is the interaction energy (in kcal mol⁻¹) between the side chains of residues i and j according to the CHARMM19 force field. Residue numbers are indicated along each of the axes.

Methods

Materials

^{15}N and ^{13}C -labelled mutants, G48M and G48V, were prepared as described previously²². Samples were 1.0 mM in protein, 50 mM sodium phosphate, 0.05% Na₂S₂O₅, 0.2 mM EDTA, pH 7.

NMR methods

Chemical shifts of the mutant proteins were assigned by using standard triple-resonance NMR techniques²³ and were subsequently used to establish that each mutant had the same secondary structure as the wild type, by using the chemical-shift index²⁴. Relaxation dispersion spectra were recorded at 600 and 800 MHz for G48V at temperatures of 10–25 °C in 5 °C intervals by methods that have been described previously^{15,25}. Dispersion spectra of G48M were obtained at 15–35 °C in 5 °C steps at 500, 600 and 800 MHz, except at 15 °C where fields of 600 and 800 MHz only were employed. Three duplicate points for each dispersion were recorded for error analysis. Dispersion profiles for each residue were considered for subsequent computations with a set of criteria described in Supplementary Information. Dispersion curves were fitted analytically with either two-site or three-site exchange models by using software written in house. For each mutant, all dispersion profiles were fitted together, enforcing the relation $k = (k_{\text{B}}T/h)\exp(-\Delta G^\ddagger/RT)$, where k is a rate constant, k_{B} is the Boltzmann (Planck) constant and $\Delta G^\ddagger$ is the activation free energy. Note that the values for ΔG , ΔH and ΔS are independent of the prefactor used in the expression for k . Extensive minimizations with different initial conditions were performed to ensure that minimum-energy solutions to the fits of dispersion data were obtained. Extracted parameters from fits to the model $U \leftrightarrow I \leftrightarrow F$ are k_{UI} , k_{IU} , k_{IF} and k_{FI} (at each temperature, from which populations can be calculated), chemical-shift differences between each pair of states (assumed to be temperature invariant¹⁴) and values of R_2^* extrapolated to infinite CPMG repetition rate for each amide ^{15}N , at each spectrometer field and temperature. For all initial fits (and results in Fig. 2) ΔC_p values were set to 0, so that the extrapolated thermodynamic parameters correspond approximately to those at the temperature midpoint of the studies. The ΔG values extracted in this way are essentially independent of ΔC_p for the range of temperatures studied owing to the compensation between ΔH and ΔS . Additional fits were performed assuming $\Delta C_p = 300 \text{ cal mol}^{-1} \text{ K}^{-1}$, estimated from the slope of the plot of ΔH against T_m obtained for a large number of G48 mutants¹³. Here we assumed that the transition state between U and I, IU, and all subsequent states on the folding pathway have similar levels of compactness so that ΔC_p is attributed to differences between U and IU. Qualitatively, the energy profiles of Fig. 2c are very similar (details are given in Supplementary Table 3S).

Concentration-dependent studies have been performed on the G48M mutant, establishing that the presence of the third state in the exchange processes is not due to oligomerization (see Supplementary Fig. 2S).

Structure calculations

^{15}N chemical shifts were used as restraints for the calculation of ensembles representing the I state for G48M and G48V, with the use of an approach developed for the calculation of transition-state structures from protein engineering ϕ values, described previously^{19,26}. The exact definition of native contact, the expression used to compute $N_i(k)$ and that used for the restraint violation term can be found in Supplementary Information, as well as a detailed description of the protocol used for the structure determination.

Received 8 January; accepted 17 May 2004; doi:10.1038/nature02655.

1. Fersht, A. *Structure and Mechanism in Protein Science* (W.H. Freeman and Company, New York, 1999).
2. Kahn, F., Chuang, J. I., Gianni, S. & Fersht, A. R. The kinetic pathway of folding of barnase. *J. Mol. Biol.* **333**, 169–186 (2003).
3. Sanchez, I. E. & Keifhaber, T. Hammond behavior versus ground state effects in protein folding: Evidence for narrow free energy barriers and residual structure in unfolded states. *J. Mol. Biol.* **327**, 867–884 (2003).
4. Klein-Seetharaman, J. et al. Long-range interactions within a nonnative protein. *Science* **295**, 1719–1722 (2002).
5. Pawson, T. & Gish, G. D. SH2 and SH3 domains: from structure to function. *Cell* **71**, 359–362 (1992).

6. Jackson, S. E. How do small, single domain proteins fold? *Fold. Design* **3**, R81–R91 (1998).
7. Dobson, C. M. Protein folding and misfolding. *Nature* **426**, 884–890 (2003).
8. Daggett, V. & Fersht, A. R. The present view of the mechanism of protein folding. *Nature Rev. Mol. Cell Biol.* **4**, 497–502 (2003).
9. Northey, J. G. B., Di Nardo, A. A. & Davidson, A. R. Hydrophobic core packing in the SH3 domain folding transition state. *Nature Struct. Biol.* **9**, 126–130 (2002).
10. Plaxco, K. W. et al. The folding kinetics and thermodynamics of the Fyn-SH3 domain. *Biochemistry* **37**, 2529–2537 (1998).
11. Martinez, J. C. & Serrano, L. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010–1016 (1999).
12. Grantcharova, V. P., Riddle, D. S. & Baker, D. Long-range order in the src SH3 folding transition state. *Proc. Natl Acad. Sci. USA* **97**, 7084–7089 (2000).
13. Di Nardo, A. A., Korzhnev, D. M., Stogios, P. J., Zarrine-Afsar, A., Kay, L. E. & Davidson, A. R. Dramatic acceleration of protein folding by stabilization of a non-native backbone conformation. *Proc. Natl Acad. Sci. USA* **101**, 7954–7959 (2004).
14. Palmer, A. G., Kroenke, C. D. & Loria, J. P. NMR methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods Enzymol.* **339**, 204–238 (2001).
15. Mulder, F. A. A., Mittermaier, A., Hon, B., Dahlquist, F. W. & Kay, L. E. Studying excited states of protein by NMR spectroscopy. *Nature Struct. Biol.* **8**, 932–935 (2001).
16. Dunitz, J. D. Win some, lose some. Enthalpy–entropy compensation in weak intermolecular interactions. *Chem. Biol.* **2**, 709–712 (1995).
17. Oliveberg, M., Tan, Y. J. & Fersht, A. R. Negative activation enthalpies on the kinetics of protein folding. *Proc. Natl Acad. Sci. USA* **92**, 8926–8929 (1995).
18. Dobson, C. M., Sali, A. & Karplus, M. Protein folding: A perspective from theory and experiment. *Angew. Chem. Int. Edn Engl.* **37**, 867–893 (1998).
19. Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. Three key residues from a critical contact network in a protein folding transition state. *Nature* **409**, 641–645 (2001).
20. Xu, X. P. & Case, D. A. Probing multiple effects on ^{15}N , $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, $^{13}\text{C}'$ chemical shifts in peptides using density functional theory. *Biopolymers* **65**, 408–423 (2002).
21. Northey, J. G. B., Maxwell, K. L. & Davidson, A. R. Probing folding kinetics beyond the Phi value: Using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state. *J. Mol. Biol.* **320**, 389–402 (2002).
22. Maxwell, K. L. & Davidson, A. R. Mutagenesis of a buried polar interaction in an SH3 domain: sequence conservation provides the best prediction of stability effects. *Biochemistry* **37**, 16172–16182 (1998).
23. Bax, A. Multidimensional nuclear magnetic resonance methods for protein studies. *Curr. Opin. Struct. Biol.* **4**, 738–744 (1994).
24. Wishart, D. S. & Sykes, B. D. The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J. Biomol. NMR* **4**, 171–180 (1994).
25. Tollinger, M., Skrynnikov, N. R., Mulder, F. A. A., Forman-Kay, J. D. & Kay, L. E. Slow dynamics in folded and unfolded states of an SH3 domain. *J. Am. Chem. Soc.* **123**, 11341–11352 (2001).
26. Paci, E., Vendruscolo, M., Dobson, C. M. & Karplus, M. Determination of a transition state at atomic resolution from protein engineering data. *J. Mol. Biol.* **324**, 151–163 (2002).
27. Noble, M. E., Musacchio, A., Saraste, M., Courtneidge, S. A. & Wierenga, R. K. Crystal structure of the SH3 domain in human Fyn; comparison of the three-dimensional structures of SH3 domains in tyrosine kinases and spectrin. *EMBO J.* **12**, 2617–2624 (1993).
28. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. *Numerical Recipes in C* (Cambridge Univ. Press, 1988).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported through grants from the Canadian Institutes of Health Research (A.R.D., L.E.K. and D.M.K.), the European Commission (X.S.), the Royal Society (M.V.), the Leverhulme Trust (M.V. and C.M.D.) and the Wellcome Trust (C.M.D.). L.E.K. holds a Canada Research Chair in Biochemistry.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.E.K. (kay@pound.med.utoronto.ca) or C.M.D. (cmd44@cam.ac.uk).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayouni (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequinot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Haraikatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejpn.com

naturejobs

Up for review

To many students attempting to get onto a graduate programme or to secure a postdoctoral fellowship, the idea of having to undergo peer review for such positions may sound extreme. But a recent study suggests that this could provide a good indication of future success — and might help to eliminate unintentional bias along the lines of gender or nationality.

The Foundation for Basic Research in Biomedicine in Heidesheim, Germany, uses peer review to assess applicants for its PhD scholarships. Social-science researchers at the Swiss Federal Institute of Technology in Zurich analysed nearly 2,700 of these scholarship applications submitted between 1985 and 2000, and found that more than 98% of the fellowship holders successfully completed their PhDs (L. Bornmann and H.-D. Daniel *B.I.F. FUTURA* 19, 7–19 (2004); www.bifonds.de/public/news/bornmann_e.pdf). In addition, the scientific papers produced by the fellows were frequently published in high-impact journals and had higher than average citation figures. The researchers also noted that neither nationality nor gender seemed to play a role in the selection process for the fellowship.

Using peer review to assess prospective graduate students and postdocs might be a way to bypass problems such as cronyism, especially in countries that rely more on patronage and connections than on qualifications. And for European Union programmes, peer review might be a better criterion for recommending funding for young scientists than amorphous concepts such as potential contributions to undefined scientific networks.

The biggest concern with expanded peer review has to do with potential. Subjecting some postgraduates to rigorous review before they have received training means that they might not have the chance to realize their promise, even if review is a good indicator of others' future success.

Paul Smaglik
Naturejobs editor



Contents

CAREER VIEW

Nuts & Bolts
Seeking feedback
Graduate Journal
The 'lab widow'
Movers
Charles Alcock

p592

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

NUTS & BOLTS

GRADUATE JOURNAL

The 'lab widow'

It's 6 p.m.. I'm in the lab and my wife rings: "When will you be home?"

"Well," I say, "I've got to finish some mini-preps, plate out some bacteria, set up a PCR and transfect some cells, so I should be there in a couple of hours. Maybe around 8 p.m. or so."

At 10.15 p.m. I call my wife. She doesn't answer (probably annoyed that I'm so late). I leave a message on the answering machine and rush home.

After regularly repeating this episode, I have concluded that I am perhaps the worst-ever judge of 'lab time'. I'm so bad that my wife automatically increases my estimated arrival time by a third. She has a 'normal' job with an eight-hour day, whereas I'm doing crazy trying-to-finish-my-thesis hours, which run from about 10 a.m. until whenever, every day. At times, we can go for days seeing each other only in passing.

When she told me that she was beginning to feel like a 'lab widow', I knew I had to do something. So now, I try to reserve at least one night every week for us.

It is hard for scientists to maintain relationships with partners who don't work in the lab. Maybe that is why inter- and intra-lab romances are so common. Passion for one's research is a good thing. But we should try to be at least equally passionate about those who have passion for us. ■

Tshaka Cunningham is a fifth-year graduate student at Rockefeller University in New York.

Seeking feedback

Most people relegate professional feedback to an annual review or an approaching tenure decision. Some have such an aversion to criticism that they try to avoid it altogether. Whatever your approach has been, consider the benefits of seeking feedback regularly and informally.

Feedback can affirm your self-awareness, identify skills gaps and align priorities. At the very least, it can reduce the fear of the unknown. Rather than restricting feedback to a formal process, seek it regularly from a variety of sources. To get the most from it, focus on three key elements: who to ask, how to receive it and what to do with it.

The people to ask for feedback are individuals who are familiar with your work, have credibility in your eyes and who will be honest with you. An



With Deb Koen
Careers consultant

effective feedback provider will be specific, offer examples, and address strengths as well as potential problems.

Select one person to start with and set up a meeting time. During the discussion, listen carefully to comments without interrupting. Don't resort to denying, defending, attacking or withdrawing from unfavourable views. Seek clarification and ask for specific examples if you're uncertain about the message. For feedback on your work, consider the following questions. What strengths do you feel I have?

What is my reputation? Are there any tough messages I need to hear? What talents am I not using? What actions would you suggest that I take? What other people or resources would you recommend to further my development?

Feedback comes to life only after the discussion, when you consider its implications and make choices about your follow-up. Document the discussion for future reference, and look for patterns in responses, giving particular attention to themes that emerged from more than one source. Finally, create a development plan for your own use or for integration into the existing review process.

By doing this, the feedback process will greatly enhance your personal and professional development. ■

Deb Koen is vice-president of Career Development Services and a columnist for *The Wall Street Journal's CareerJournal.com*.

MOVERS

Charles Alcock, director, Harvard-Smithsonian Center for Astrophysics, Massachusetts



Charles Alcock gained autonomy relatively early in his career — although it proved to be something of a mixed blessing. As a fellow at Princeton's Institute for Advanced Study in New Jersey, he had a respectable stipend and stimulating weekly lunches with other astrophysicists, but he didn't have to do any teaching, grant-writing or administration.

The downside to this was that he quickly had to learn those additional skills during his first professorship at the Massachusetts Institute of Technology (MIT) in Cambridge. "The life of a professor is demanding," he says.

But Alcock was pleased to find that responsibility outside research can bring its own rewards. Teaching a course in statistical mechanics at MIT allowed him to "correct a gap in my own education", he says. And the "outstanding" standard of students there forced him to prepare rigorously for his lectures.

If his time at MIT helped him with his teaching, Alcock learned most of his management skills in his next post at the Lawrence Livermore National Laboratory in California. Alcock says he was lucky that his first boss at Livermore was Claire Max — now professor of astronomy and astrophysics at the University of California, Santa Cruz — who had an almost intuitive

understanding of how to get people to work together on large-scale projects.

"She had a very good sense of how to ask other people what they could do," Alcock says. "She was very interested in other peoples' skills and finding out what they were interested in."

With Max, Alcock learned how to manage his own large project, the Massive Compact Halo Object (MACHO), which aims to monitor the brightness of 1.8 million stars. "I don't think I had any of the skills needed to get a group of 18 or 20 people working together," Alcock says.

At his new position, heading the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, that skill will be even more valuable. The centre has more than 300 PhD scientists associated with it, and several large-scale projects running simultaneously. But Alcock believes he can show that his early autonomy is no barrier to success. ■

CV

2000–04: Professor of astronomy and astrophysics, University of Pennsylvania, Philadelphia

1986–2000: Director, Institute of Geophysics and Planetary Physics, Lawrence Livermore National Laboratory, California

1981–86: Associate professor of physics, Massachusetts Institute of Technology, Cambridge

1977–81: Long-term member, Institute for Advanced Study, Princeton, New Jersey

1977: PhD in astronomy and physics, California Institute of Technology